ROBUST BINAURAL NOISE-REDUCTION STRATEGIES WITH BINAURAL-HEARING-AID CONSTRAINTS: DESIGN, ANALYSIS AND PRACTICAL CONSIDERATIONS

A Dissertation
Presented to
The Academic Faculty

by

Jorge I. Marin

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Electrical and Computer Engineering

Georgia Institute of Technology
August 2012

ROBUST BINAURAL NOISE-REDUCTION STRATEGIES WITH BINAURAL-HEARING-AID CONSTRAINTS: DESIGN, ANALYSIS AND PRACTICAL CONSIDERATIONS

Approved by:

Professor David V. Anderson, Advisor
School of Electrical and Computer Engineering
Georgia Institute of Technology

Professor Pamela T. Bhatti
School of Electrical and Computer Engineering
Georgia Institute of Technology

Professor Mark A. Clements
School of Electrical and Computer Engineering
Georgia Institute of Technology

Professor Gil Weinberg
College of Architecture
Georgia Institute of Technology

Professor Christopher Rozell
School of Electrical and Computer Engineering
Georgia Institute of Technology

Date Approved: 14 May 2012

To my parents

ACKNOWLEDGEMENTS

I want to thank my advisor, my committee, and all my colleagues at Georgia Tech who made my academic and professional growth possible. To my family for their support during the time I was abscent from my home country. To the Universidad del Quindio–Colombia for its financial support and to allow me to pursue my PhD studies. To the Fulbright program and Colciencias–Colombia for its financial support during the first stage of my PhD program. To National Semiconductor Corporation (now Texas Instruments Inc.) and Starkey Laboratories for their contribution to my profesional growth as well as their partial support to this research.

Contents

List of Tables

List of Figures

LIST OF SYMBOLS OR ABBREVIATIONS

BSS Blind Source Separation.

BSS-PP Blind Source Separation with Perceptual Post Processing.

CPSD Cross Power Spectral Density.

DB-MWF Distributive Binaural MWF.

DoA Direction of Arrival.

DSP Digital Signal Processor.

DWT Discrete Wavelet Transform.

DWT-PMWF Discrete Wavelet Transform PMWF.

FB Filterbank–Term referred to an IIR filterbank that resembles the auditory filterbank.

FB-PMWF IIR Filter Bank PMWF.

FFT Fast Fourier Transform.

FFT-MWF Term used referred to the SDW-MWF method implemented with FFT processing.

FW Frequency-Warped Filter.

FW-PMWF Frequency-Warped PMWF.

GSC Generalized Sidelobe Canceller.

HRTF Head-Related Transfer Function.

I3 Objective Intelligibility Metric.

IDBM Ideal Binary Masking.

ILD Interaural Level Difference.

ITD Interaural Time Difference.

MMSE Minimum Mean-Square Error.

MVDR Minimum-Variance Distortionless Response.

MWF Multichannel Wiener Filter.

MWF-CPSD MWF method that uses CPSD estimators to obtain the speech and noise correlation matrices.

MWF-IDBM MWF with Post Processing Using Ideal Binary Masking.

MWF-N MWF with partial noise.

NPLR Noise Power Level Reduction.

PEMO-Q Perceptual Model of Quality.

PESQ Perceptual Evaluation of Speech Quality.

PMWF Perceptually-Inspired MWF.

PSD Power Spectral Density.

RECUP-MWF Recursive Update MWF.

SDW-MWF Speech Distortion Weighted MWF.

SNR-SII Broadband Intelligibility-Weighted SNR.

VAD Voice Activity Detector.

WDFT Warped Discrete Fourier Transform.

WP Wavelet Packet.

WP-PMWF Wavelet Packet PMWF.

SUMMARY

The objective of the dissertation research is to investigate noise reduction methods for binaural hearing aids based on array and statistical signal processing and inspired by a human auditory model. In digital hearing aids, wide dynamic range compression (WDRC) is the most successful technique to deal with monaural hearing losses. This WDRC processing is usually performed after a monaural noise reduction algorithm. When hearing losses are present in both ears, i.e., a binaural hearing loss, independent monaural hearing aids have been shown not to be comfortable for most users, preferring a processing that involves synchronization between both hearing devices. In addition, psycho-acoustical studies have identified that under hostile environments, e.g., babble noise at very low SNR conditions, users prefer to use linear amplification rather than WDRC. In this sense, the noise reduction algorithm becomes an important component of a digital hearing aid to provide improvement in speech intelligibility and user comfort. Including a wireless link between both hearing aids offers new ways to implement more efficient methods to reduce the background noise and coordinate processing for the two ears. This approach, called binaural hearing aid, has been recently introduced in some commercial products but using very simple processing strategies. This research analyzes the existing binaural noise-reduction techniques, proposes novel perceptually-inspired methods based on blind source separation (BSS) and multichannel Wiener filter (MWF), and identifies different strategies for the real-time implementation of these methods. The proposed methods perform efficient spatial filtering, improve SNR and speech intelligibility, minimize block processing artifacts, and can be implemented in low-power architectures.

Chapter I

INTRODUCTION

Digital hearing aids are recognized as an efficient way to aid people with mild to severe hearing losses. For these hearing losses, a dynamic amplification of the audio signals coming into the ear is performed to overcome the cochlear damage. A digital hearing aid is usually composed of five functional blocks: Directional processing, noise reduction, compression, feedback cancellation, and sound classification. The purpose of directional processing and noise-reduction blocks is to enhance the target signal. In the compression block, the particular information about the user's hearing loss is used to perform an amplification of the enhanced signal. In some hearing aids, a feedback cancellation block is required to avoid over-amplification due to the presence of an acoustic feedback path. Finally, the signal classification block is used to detect the features of the target signal (speech or music) and the environmental condition (quiet or noisy place) to set up the parameters of the other functional blocks.

Hearing losses can be present in both ears, which is called a binaural hearing loss. Although independent monaural hearing aids have been traditionally used to deal with binaural hearing losses, recent research have been shown that binaural processing, i.e., processing that mixes the information of both ears, provides significant advantages over monaural processing.

Two functional blocks may be improved with binaural processing: Compression and noise reduction. Although different binaural compression algorithms have been proposed in the literature [33], psycho-acoustical studies have identified that under adverse environments, users prefer linear amplification over compression [24, 66]. Hence, the benefits of binaural compression are insignificant for these environments. On the contrary, the benefits of binaural noise reduction are significant for these environments to improve user comfort and speech intelligibility.

Power consumption and computational resources are two important issues and challenges for the design of a digital hearing aid. Existing binaural noise-reduction algorithms proposed in the literature are robust against highly noisy environments but require complex and sophisticated digital signal processors (DSP) for their implementation. Hence, the development of commercial binaural hearing aids has been limited to simple algorithms whose performance is not as effective as that of the sophisticated methods [70, 73, 87].

The objective of the proposed research is to investigate noise-reduction methods for binaural hearing aids based on array and statistical signal processing and inspired by a human auditory model. Most binaural noise-reduction algorithms proposed in the literature are designed to meet certain objective metrics, e.g. minimization of the mean-square error between the desired output and the system output among other metrics. These approaches do not provide an easy and efficient way to satisfy the implementation constraints imposed by binaural-hearing-aid devices. However, we know from studies in human perception that under certain conditions some information may be redundant or unnecessary. Hence, we believe that by using a human auditory model to drive design decisions, it will be possible to achieve improvements in quality, computation costs, and to meet the hardware and other implementation constraints of a binaural hearing aid. Including a human auditory model in the design of the noise-reduction methods may make possible for the resulting algorithms to show improvement in noise reduction, speech quality, and speech intelligibility. In addition, these auditory models may allow us to remove redundant information from the perceptual viewpoint, obtaining feasible implementations for the algorithms by reducing the computational cost and wireless transmission bandwidth. These proposed methods should perform efficient spatial filtering, improve speech intelligibility, minimize block processing artifacts, and be implementable in low-power architectures such as binaural hearing aids. To accomplish these goals, this document presents the study of existing binaural noise-reduction techniques, the design of novel robust binaural noise-reduction methods, and identifies different strategies for the real-time implementation of these techniques. The novel binaural noise-reduction methods proposed in this research are based on blind source separation (BSS) and multichannel Wiener filters (MWF), employing strategies inspired by

a human auditory model.

This dissertation is organized as follows. Existing binaural noise-reduction methods are surveyed in Chapter 2. This chapter also discusses the challenges and open problems to implement noise-reduction algorithms in binaural hearing-aid devices. Next, Chapter 3 presents a comparative study about the existing binaural noise-reduction methods and identifies the promising techniques to implement a binaural hearing aid. These promising techniques are based on BSS and MWF, and improvements on these techniques will be the focus on the next two chapters. Chapter 4 introduces a novel binaural noise-reduction strategy based on blind source separation (BSS) and perceptual post processing (BSS-PP), and Chapter 5 describes different binaural noise-reduction strategies based on MWF using perceptual processing (PMWF). The performance of the proposed methods, BSS-PP and PMWF, is discussed on the Chapter 6. Additional implementation strategies are described in the Chapter 7. Finally, the contributions and future research is presented in Chapter 8.

Chapter II

OVERVIEW OF BINAURAL NOISE-REDUCTION METHODS

This chapter presents an overview of the existing binaural noise-reduction algorithms and the challenges to implement these algorithms on real devices.

2.1 Binaural Processing vs. Monaural Processing

Most hearing-impaired people suffer from hearing losses at both ears, called binaural hearing loss. Traditionally, independent monaural hearing aids have been used to deal with binaural hearing losses. However, recent research on hearing aids have been shown that binaural processing is a promising field to achieve more aggressive goals in terms of directional noise cancellation, improvement in speech intelligibility under hostile environments (e.g., babble noise), and user comfort [76, 95, 89, 90]. Binaural processing refers to the processing that mixes the information received at the microphones of both hearing aids to finally deliver enhanced signals to each ear. In other words, it can be seen as a multiple-input, two-output system that includes a communication link between both hearing devices.

The binaural processing techniques reported in the literature comprise two categories: Compression and noise reduction. Although different binaural compression methods have been reported, their performance is similar to those of monaural compression methods [33]. On the contrary, the benefits provided by binaural noise-reduction methods have been shown to be very significant compared to those of monaural methods. This fact has been explained through psycho-acoustical studies about the hearing perception in hostile environments (e.g., babble noise at very low SNR conditions). These studies show a user preference for linear amplification over compression [24], and the relevance of the preservation of localization cues[1] for target identification and speech intelligibility [76, 89, 90]. Preserving

[1]Localization cues are also known as binaural cues in some reports.

localization cues provides information about the direction of arrival of the target and interfering signals, and this information may aid the noise separation performed by the human auditory system. For independent monaural hearing aids, the lack of a communication link to synchronize both hearing aids may lead to a distortion or loss of localization cues, which has been recognized as perceptually annoying [99, 64, 89].

2.2 Noise-Reduction Methods in Binaural Hearing Aids

Some binaural noise-reduction techniques proposed in the literature enhance exclusively the target signal coming from the front. Since the preservation of localization cues is an important feature of a binaural hearing aid, this research is focused only on the techniques that enhance the target signal coming from any arbitrary direction of arrival. Existing techniques have been shown to be successful for the reduction of interfering signals in multi-talker, diffuse and babble noise (also known as cafeteria noise) environments; the preservation of localization cues for the target signal; and in a very few cases, the preservation of localization cues for the interfering signals. Although the processing in most binaural noise-reduction techniques involves two signals, the signals at the left and right hearing aids, some techniques support multiple microphones per hearing aid to improve the noise reduction. In the following sections, an overview of the more representative binaural noise-reduction techniques is presented. These noise-reduction techniques are grouped and discussed in four categories: Scene analysis, adaptive beamforming, multichannel Wiener filtering (MWF), and blind source separation (BSS).

2.2.1 Scene Analysis

Scene analysis approaches are characterized by the use of measurements taken from the input signal to compute a set of frequency responses that are used to filter out the noise. Former scene analysis approaches assume that the target signal is coming from the front [102]. To avoid this limitation, alternative methods allowing any arbitrary direction of arrival (DoA) of the target signal are proposed in [3, 42]. In [3], Chisaki *et al.* proposed a method that measures the interaural time difference (ITD) and the interaural level difference (ILD) of the input signals. Using these measurements, the DoA of the target signal is

estimated by searching in a database, then the head related transfer function (HRTF) corresponding to the estimated DoA is used as frequency response to perform the filtering. Although this approach is inspired by a perceptual model of the binaural hearing, the authors detected that the relationship between the ILD and DoA is not unique, leading to an ambiguous DoA estimation. On the other hand, in the method proposed by Li *et al.* [42], a noise signal is estimated using the information from the left and right microphones, and this noise signal is subtracted from each input to get the enhanced signals for the left and right hearing aids. This method assumes that the target signal is in-phase at both sides, which is not true for all frequencies and DoAs due to the head shadow effect.

Recent scene analysis techniques that exploit the properties of the coherence function between the left and right signals [31, 74] have become more successful than the techniques discussed above. In these methods, the coherence function is used to estimate a frequency response to cancel out the background noise and interferences. Kamkar *et al.* [31] proposed a method derived from the coherence function to estimate the power spectrum density (PSD) of the noise, and using this noise PSD, to perform a speech enhancement. On the other hand, Rahmani *et al.* [74] described an alternative method that uses the noise cross-PSD (CPSD) instead of the noise PSD as in Kamkar's method. Although the Rahmani's method is focused on the noise cancellation for a hands-free kit, similar ideas may be explored for a binaural hearing aid. Both Kamkar's and Rahmani's methods are attractive because they do not require a voice activity detector (VAD).

Finally, a recent strategy based on a two-stage processing and Wiener filter (TS-BASE / WF) was proposed by Li *et al.* [43, 44]. These two stages refer to equalization and cancellation of the target signal to get a noise estimate, and using this noise estimate to compute the parameters of a Wiener filter. This technique outperforms other spectral subtraction and beamforming approaches and preserves the localization cues of the target signal.

2.2.2 Adaptive Beamforming

Since a binaural hearing aid is a sensor array, noise-reduction techniques based on array processing are suitable to implement a binaural hearing aid. A well-known array processing strategy for noise reduction is to perform beamforming at the direction of arrival (DoA) of the target signal. In addition to the knowledge of DoA for the target signal and/or interfering signals, a beamformer for binaural hearing aids requires a post processing to recover the localization cues. One of the first beamformers for binaural hearing aids was proposed by Lotter and Vary [48]. This method, called superdirective beamforming, is based on a minimum-variance distortionless response (MVDR) beamformer.

Rohdenburg *et al.* [79, 78] proposed two methods for noise cancellation in binaural hearing aids that improve the Lotter and Vary's method by allowing continuous tracking of the target signal. Rohdenburg's methods use a generalized side-lobe canceler (GSC), in which a MVDR beamformer is used for the main channel, and the DoA is estimated by an algorithm that takes into account the head-related transfer functions (HRTF) [20]. The signal at the output of the MVDR beamformer is used to estimate the frequency-domain suppression gains for both ears. These methods show good performance to reduce diffuse or ambient noise but low performance to reduce babble noise.

Rohdenburg's approaches are based on a narrow-band beamformer, but a more accurate approach to process speech signals is to use wide-band beamformers. Nishimura *et al.* [68] addressed the problem of using wide-band beamformers for binaural noise reduction. In this approach, constraints to preserve perceptual cues at both ears are introduced in a wide-band MVDR beamformer. The disadvantage of this technique is its high computational cost.

2.2.3 Multichannel Wiener Filter (MWF)

Wiener filters have been widely recognized as an effective strategy for noise reduction in stationary processes. The coefficients in a Wiener filter are estimated through the minimization of a cost function, typically the mean-square error between the desired signal (clean speech) and the Wiener filter output. In this framework, the signal and noise statistics are the only requirement to compute the filter coefficients. This feature affords robustness

to the Wiener filter because the filter is able to enhance the target signal arriving from any arbitrary direction. The multi-sensor variant of a Wiener filter is called multichannel Wiener filter (MWF), and it was initially introduced by Doclo and Moonen [11] for speech enhancement.

Following the Doclo and Moonen's ideas [11], several authors proposed the use of MWF for noise reduction in binaural hearing aids, in which the acoustic signals received at one hearing aid are fully transmitted to the contralateral hearing aid over a wireless link [14, 39]. In particular, the technique discussed in [14], called SDW-MWF (speech distortion weighted MWF), is the most widely-known MWF technique for binaural hearing aids. The benefits of using MWF for binaural noise reduction over monaural processing have been shown in [95, 97, 98]. These benefits of MWF are related to a better noise suppression and an effective preservation of the localization cues for the target signal.

Since then, alternative methods to reduce the bandwidth of the communication link has been analyzed in [13, 15, 94, 98]. For example, Van den Bogaert *et al.* [98] proposed and analyzed an extension of SDW-MWF for partial transmission of channels. Srinivasan [94] proposed a method based on psycho-acoustical studies that showed monaural noise reduction is sufficient to process high-frequency components, while binaural noise reduction works better for low-frequency components. Thus, the input signal received at each ear is filtered to extract the low- and high-frequency components, and the low-frequency component is transmitted to the contralateral hearing aid. At each hearing aid, two Wiener filters are used, a monaural Wiener filter for the high frequencies, and a binaural Wiener filter for the low frequencies. In Srinivasan's paper, a comparison with other techniques is not discussed. On the other hand, Doclo *et al.* [13, 15] proposed four different methods to reduce the bandwidth of the communication link: MWF-Contra, MWF-Front, MWF-Superd, and DB-MWF. In these approaches, acoustic signals arriving at one particular hearing aid are combined using a linear transformation to produce a single-channel signal to be transmitted to the contralateral hearing aid. Among these methods, DB-MWF was theoretically proved to converge to SDW-MWF [13].

MWF techniques can be classified in two groups depending on the preservation of noise

localization cues. All above MWF methods do not preserve noise localization cues. Noise localization cues are preserved in the MWF framework by including special terms in the cost function. These terms lead to the addition of noise to the enhanced signal and the degradation of the output SNR. However, perceptual experiments showed that preserving noise localization cues may provide information to the auditory system to allow it to successfully remove the background noise [10, 40, 97, 98]. For example, in [10], Doclo *et al.* proposed an extension of SDW-MWF to preserve noise cues. In this technique, all channels are transmitted over the communication link, and the cost function used to derive the filter weights is modified to include an extra term that involves the interaural time difference (ITD). In this sense, including the ITD is possible to preserve the directional cues for both speech and noise. Since this method involves high computational cost, a less complex solution to preserve noise localization cues is described in [40]. This method, called MWF-N, includes a trade-off parameter to control the amount of noise to be added at the output. Although this technique requires the full the transmission of channels to the contralateral hearing aid, an extension using partial transmission of channels is discussed in [98].

Van den Bogaert *et al.* [97, 98] presented a comparative analysis between SDW-MWF, MWF-N, and the adaptive directional microphone (ADM) [49]. The latter is the most widely-known technique for noise reduction in commercial monaural hearing aids. Authors showed that ADM outperforms the MWF techniques only when the target signal is coming from the front. However, MWF approaches outperform ADM when the target signal is coming from directions other than the front, which means that MWF-based techniques offer good speech localization [97, 98]. Moreover, the direction of arrival for the interfering signals, i.e. the noise localization cues, is lost in SDW-MWF and preserved in MWF-N [97]. A theoretical proof of this fact is presented in [5]. Another interesting result was found with respect to the bandwidth required for the communication link. Perpetual tests on SDW-MWF and MWF-N showed that transmitting only one channel to the contralateral hearing aid is enough to ensure good noise removal [98].

2.2.4 Blind Source Separation (BSS)

Blind source separation (BSS) methods can be seen as multiple-input multiple-output systems, in which a mixture of signals arriving at the inputs is separated (or unmixed), and then each BSS output is expected to provide an estimate of each signal in the input mixture. These methods are based on an assumption of the mutual independence of the source signals [1, 41]. When a two-output BSS algorithm is used in noise-reduction applications, one BSS output is expected to provide an estimate of the target signal, and the another output an estimate of the interfering signals. The estimate of the target signal is not perfect because of the residual background noise present at this output [96]. Hence, single-output BSS-based noise-reduction algorithms proposed in the literature [41, 69, 72] employ a post-processing stage following the BSS to enhance the output sound quality and to improve the noise reduction. Since the localization cues are lost in the enhanced output of single-output BSS-based noise-reduction methods, the post processing employed in binaural noise-reduction methods is used to recover the localization cues as well as to enhance the target signal.

Aichner *et al.* [1] proposed two methods to preserve localization cues and compared these techniques to a previous method proposed by Wehr *et al.* [101]. Aichner *et al.* concluded that the method using a BSS algorithm and adaptive filters outperforms the other methods and preserves efficiently the localization cues. However, a study conducted in this dissertation [62, 63] identified that the preservation of the localization cues in the Aichner's approach is possible only in the determined case, i.e., when the number of interfering signals is lower than the number of microphones (Section 6.2). In [75], Reindl *et al.* proposed an alternative post-processing stage based on Wiener filter to recover the localization cues. They showed the potential benefits of this approach over the Aichner's approach, mainly to deal with complex interfering signals such as babble noise, and to preserve the localization cues of both target and interfering signals.

An alternative BSS-based method for noise cancellation, discussed in [41], uses a standard BSS algorithm and a denoising algorithm instead of a post processing stage. An adaptive filter is used for the denoising algorithm, and its single output is applied to both ears, which means that this technique does not preserve localization cues.

2.2.5 Promising Binaural Noise-Reduction Methods

A summary of the binaural noise-reduction techniques discussed in the last sections is presented in Table 1. This table includes only the promising techniques to implement a binaural hearing aid. This selection was based on two criteria: The ability to enhance the target signal coming from any arbitrary direction of arrival and the preservation of target localization cues. This table includes two additional columns, the maximum number of microphones allowed per hearing aid and the preservation of noise localization cues. A question mark is included in those techniques for which the original paper does not provide information.

There are some attempts to establish a comparison among binaural noise-reduction techniques [1, 15, 97, 98], but they include a few number of techniques or are restricted to comparing techniques under the same category. Hence, this dissertation conducted a study that considers techniques from the four categories (scene analysis, adaptive beamforming, MWF, and BSS) under stationary and non-stationary environments [51], and using objective metrics exclusively. The goal of this research is to identify the promising methods to be used in a real-time binaural hearing aid. The results of this study are presented in the Chapter 3. This study concluded that the MWF-based and BSS-based methods provide significant noise reduction and acceptable sound quality under the different scenarios analyzed.

2.3 Implementation Issues on Binaural Noise-Reduction Algorithms

The implementation of any DSP algorithm is strongly related to the computational complexity. However, a real-time implementation involves different design considerations in addition to the computational complexity. In particular, for a binaural hearing aid, processing artifacts, latency, power consumption, and communication bandwidth are other main concerns.

a) Computational Complexity. The existing binaural noise-reduction algorithms reported in the literature are very sophisticated and demand complex DSP architectures, which turns out prohibitive for a binaural hearing aid. For this reason, these existing binaural noise-reduction methods have not been used in the current commercial binaural hearing

Table 1: Existing binaural noise-reduction methods.

Class	Technique	Max. mics/ hearing aid	Preserve noise cues	Ref.	Additional Comments
Scene Analysis	Kamkar-09	1	?	[31]	
Scene Analysis	Rahmani-09	1	?	[74]	Originally designed for hands-free kit.
Scene Analysis	TS-BASE/WF	1	?	[43, 44]	
Beamforming	Lotter-06	1	Yes	[48]	
Beamforming	Rohdenburg-07	Any	Yes	[79]	Highly dependent on DOA and HRTF estimation. Bad for babble noise.
Beamforming	Rohdenburg-08	Any	Yes	[78]	Same as Rohdenburg-07 but lower computational cost.
Beamforming	Nishimura-04	Any	Yes	[68]	High computational cost.
MWF	SDW-MWF	Any	No	[13, 15]	Full or partial transmission of channels.
MWF	DB-MWF	Any	No	[13, 15]	Reduced-bandwidth MWF.
MWF	MWF-Contra	Any	No	[13, 15]	Reduced-bandwidth MWF.
MWF	Doclo-05	Any	Yes	[10]	Full transmission of channels. Complex implementation.
MWF	MWF-N	Any	Yes	[40]	Full or partial transmission of channels.
MWF	Srinivasan-08	1	?	[94]	Reduced-bandwidth MWF.
BSS	Aichner-07	Any	Yes	[1]	Claimed to preserve noise cues, but showed the contrary by others.
BSS	Reindl-10	Any	Yes	[75]	

aids. Instead, commercial devices have been adopted simple strategies to synchronize both hearing devices [70, 73, 87]. Hence, this dissertation investigates the advantages, disadvantages, and computational complexity of the existing binaural noise-reduction methods to identify the more promising methods for a binaural hearing aid (Chapter 3). In addition, this research also proposes different strategies for the real-time implementation of these methods in low-power DSP architectures (Chapters 4 and 5).

b) Processing Artifacts. Hearing aids should use algorithms feasible to implement in a simple DSP architecture and use minimum memory requirements, but simplified algorithms may produce audible artifacts. The implementation of the majority of the binaural techniques referenced in the Table 1 involves multiplication in the frequency-domain, which is usually implemented using FFT-based processing. In most speech enhancement techniques based on FFT processing, the weight vector is updated typically in the frequency-domain by means of a non-linear algorithm [47], which may invalidate the condition to avoid circular convolution, and then audible artifacts may be perceived at the output [86]. In digital hearing aids, smoothing of the weight vector is typically employed to minimize processing artifacts [33]. A detailed mathematical analysis to identify the source of the block-processing artifacts and distortions in the weight vector was conducted in this dissertation and reported in [52, 53]. A conclusion derived from our analysis is that windowing and overlapping does not ensure the fulfillment of the condition to avoid circular convolution [52, 53]. In addition to the mathematical analysis of the FFT-based block processing artifacts, this research also proposes two processing methods to avoid these artifacts (Section 7.3). To avoid processing artifacts, noise-reduction algorithms using perfect reconstruction processing or auditory filterbanks are other desirable methods for hearing aids. This research explores these alternatives in the Chapters 4 and 5.

c) Latency. Digital hearing aids require minimum latency to avoid the delivery of unpleasant sounds to the ear. The perception of these latencies is frequency-dependent [33]. latencies larger than 2 ms are perceptible for clicks and short-time speech segments. For long-time speech segments, larger latencies are allowed, but delays larger than 15 ms are

annoying. Most binaural noise-reduction techniques require block processing implementation, which carries out a large latency. Although the latency may be reduced using shorter blocks, the frequency resolution and performance of these techniques is highly compromised [13]. So far, there is not a consensus about the most suitable way to reduce latency in block processing implementations without increasing computational complexity. However, techniques inspired by an auditory-domain processing have been shown to be promising to reduce latency in wide dynamic range compression (WDRC) [35]. This fact motivates the exploration of such methods to implement binaural noise-reduction methods.

d) Power Consumption and Bandwidth of the Wireless Link. The reduction of power consumption is related to the reduction in the computational complexity. However, a binaural hearing aid offers alternative ways to achieve this goal. Because of the existence of a wireless link, a careful attention should be focused on this component. Reducing the number of transmitted channels and reducing the bandwidth of the communication link are two ways to save power on a binaural hearing aid. These issues have been extensively explored for MWF techniques, concluding that the transmission of a single channel [98] or the use of distributive techniques [13, 83] provides similar performance to those methods employing a full transmission of channels. This transmission bandwidth may be reduced by using a suitable channel codification, but the codification itself may increase the computational complexity. To avoid a sophisticated channel codification, this dissertation proposes a method that reduces computational complexity and transmission bandwidth simultaneously. This method is discussed in the Chapter 5.

Another strategy to reduce the power consumption in binaural hearing aids may be the analysis of the current environmental condition, and depending on this condition apply a monaural or a binaural processing. In this sense, the communication link might be powered down for those situations in which binaural processing is optional (i.e., scenarios where binaural processing offers equal or lower performance than a monaural processing). An attempt to answer this question was conducted by Srinivasan [95], who studied the MWF-N method under a single-interfering-signal scenario. However, a more exhaustive analysis considering several kind of scenarios and techniques has not been published yet.

e) Parameter Estimation. There are additional implementation issues that must be addressed to implement the techniques reported in the literature. Although the selection of parameters for a real-time implementation of the scene analysis and the BSS techniques indicated in Table 1 is described completely in the original papers, this issue is not completely clear for the beamforming and MWF techniques. Beamforming techniques are highly dependent on the propagation model and DoA estimation. Rohdenburg *et al.* [78] discusses the effect of different propagation models and DoA algorithms for the binaural noise-reduction techniques proposed by them [20, 78, 79]. Hence, this analysis should be extended to the other beamforming techniques. On the other hand, most reports on MWF techniques use a voice activity detector (VAD) to update the statistics for the noise and signal. In this context, noise statistics are updated during noise-only segments, and signal statistics during voiced segments. In practical applications, using a real VAD conveys different challenges for highly-noisy and non-stationary environments. Doclo *et al.* [13] showed that the performance of SDW-MWF is degraded when the VAD errors are greater than 20%. In their analysis, authors assumed stationary background noise, but in real situations, the subject is exposed to non-stationary environments. Even using a perfect VAD, a VAD-based framework is unable track efficiently the statistics of non-stationary environments because noise statistics cannot be updated during voiced segments. Recent attempts to protect SDW-MWF from VAD errors has been introduced in [7]. Furthermore, the estimation of the statistics is not the only relevant issue to implement a MWF technique. MWF techniques include a trade-off parameter, which is usually fixed. A theoretical study showing the selection of this trade-off parameter under stationary environments for SDW-MWF and MWF-N is discussed in [5]. Other authors have proposed an adaptive trade-off parameter based on a soft VAD [67]. To track efficiently the statistics under non-stationary environments, this dissertation proposes a non-VAD second-order statistics estimation method for MWF. This method is discussed on the Section 5.3.

Chapter III

COMPARATIVE STUDY OF THE EXISTING BINAURAL NOISE-REDUCTION METHODS

Different binaural noise-reduction techniques belonging to four categories (scene analysis, adaptive beamforming, MWF, and BSS) were discussed in Section 2.2, and a list of the promising techniques was summarized in the Table 1. Since previous studies compare few methods or only methods under the same category, this research includes an extensive comparative study, employing objective metrics, to identify promising noise-reduction methods for a real-time binaural hearing aid. All techniques listed on Table 1 were implemented in Matlab, except Rohdenburg-07 [79], Nishimura-04 [68], and Doclo-05 [10] because of their computational complexity. To get a reliable identification of the promising techniques, all methods are compared assuming the best case scenario, i.e., assuming the upper-bound performance. This upper-bound performance is obtained using perfect knowledge about the DoA of the target signal and perfect VAD to estimate the parameters required by each algorithm.

Most techniques require estimates for some statistical properties, e.g., correlation matrices or power spectral densities. To simulate the effect of a real-time implementation, these statistics are updated in the frame-by-frame basis using a first-order estimator:

$$\mathbf{\Gamma}(k) = \alpha\mathbf{\Gamma}(k-1) + (1-\alpha)\gamma_k\,, \tag{1}$$

where $\mathbf{\Gamma}(k)$ is the estimate of the statistical property at the frame index k, γ_k is the instantaneous estimator of that property, e.g., $\gamma_k = \boldsymbol{x}_k\boldsymbol{x}_k^H$ for the correlation matrix, and α is a time constant that controls the smoothing of the estimator.

3.1 Experiment

We conducted simulations to discern the performance of these techniques under six different scenarios [51]:

1. Single source under constant-SNR diffusive noise. This scenario is widely used to test virtually all binaural noise-reduction techniques. This background noise is generated by playing uncorrelated pink noise sources simultaneously at 18 different spatial locations.

2. Single source under babble (or cafeteria) noise. The background noise corresponds to a real recording in a cafeteria. Although this scenario is also classified as diffusive noise, it differs from Scenario 1 in the use of a real recording. To make the distinction between Scenario 1 and 2 in this document, the term diffusive noise is used to refer to Scenario 1 and babble noise for Scenario 2.

3. Multi-talker. In this scenario, four distinguishable speakers are placed at different azimuthal positions: 40^o, 80^o, 200^o and 260^o. This scenario simulates the conditions inside an office.

4. Single source under variant-SNR background noise. The background noise in this scenario is babble noise, whose envelope is modified using two shapes. The first shape assumes that the background noise level is increasing, and the second one that the background noise level is decreasing. This non-stationary scenario simulates the effect of getting in and getting out a cafeteria.

5. Moving source in a clear environment. This simulation verifies the ability of the algorithm to track the source signal under high SNR conditions.

6. Moving source in a noisy environment. This scenario assumes the same moving pattern as Scenario 5 and constant-SNR babble noise.

The above scenarios are generated by filtering the target signal with the HRTF measured for a KEMAR manikin in absence of reverberation [19]. The target signal is placed at eight different azimuthal angles: 0^o, 30^o, 90^o, 120^o, 180^o, 240^o, 270^o and 330^o, where 0^o corresponds to the front of the KEMAR, 90^o corresponds to the right ear, and 270^o to the left ear. Target signals are speech recordings of ten different speakers and sentences taken

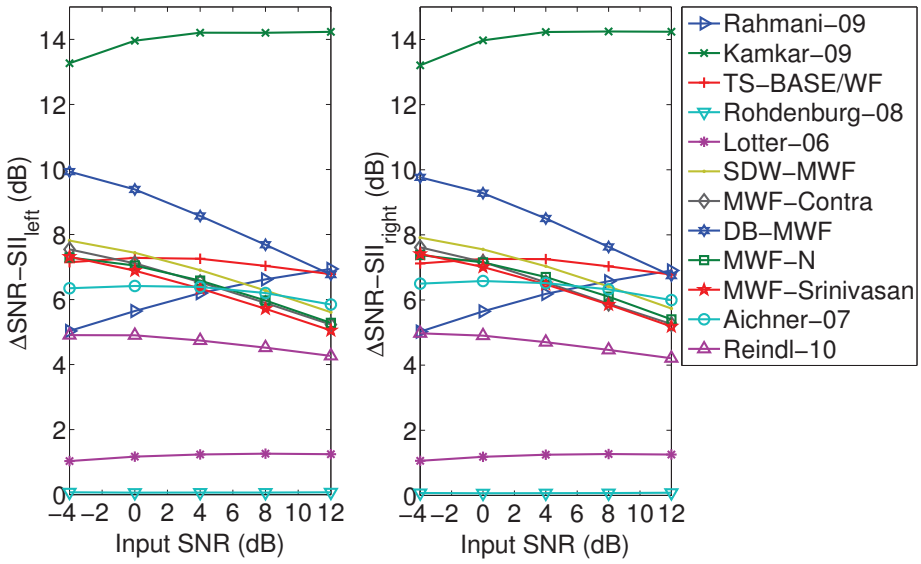

Figure 1: SNR improvement for existing techniques under diffusive noise scenario.

from the IEEE sentence database [25]. For all scenarios, the interfering signals are added to the target signal at different SNR.

The performance of these techniques is analyzed using two objective metrics, the broadband intelligibility-weighted SNR improvement (ΔSNR-SII) [21], and the objective quality assessment measure (PEMO-Q) [23]. PEMO-Q metric is selected over other objective quality assessment measurements because it showed to predict the effect of non-linear distortions more accurately than other metrics such as PEAQ [23].

3.2 Results and Discussion

Figures 1-5 show the ΔSNR-SII values obtained for all techniques under Scenarios 1, 2, 3, 4, and 6, and Figure 6 shows the PEMO-Q evaluation. Results for Scenario 5 are not included because they are used only to verify the ability of the techniques to track the non-stationary features of the signal. All methods show to be able to track the target signal. The following conclusions are derived from these figures:

- Among all techniques, Kamkar-09, a scene-analysis technique, provides the best SNR improvement (6-15dB) but the poorest output sound quality. This fact is explained through the high degree of distortion introduced by this algorithm.

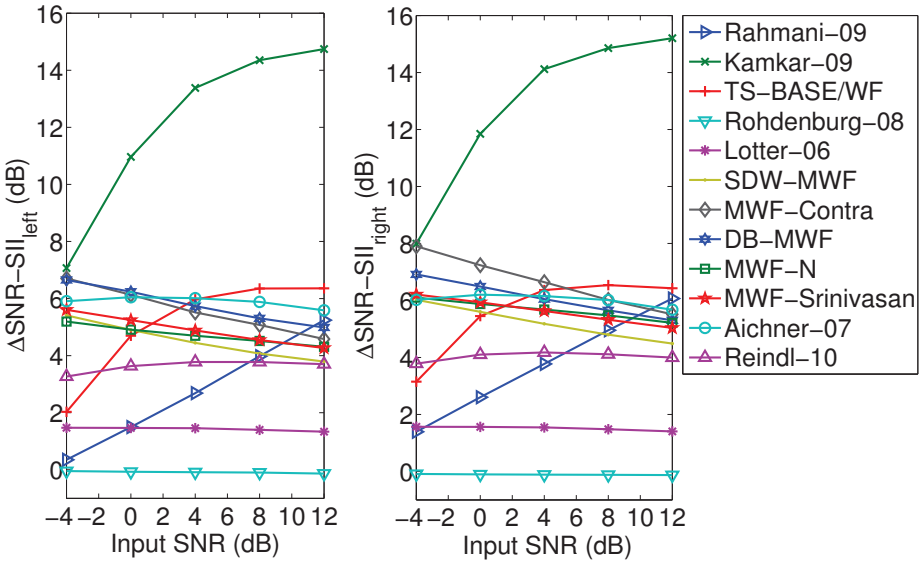

Figure 2: SNR improvement for existing techniques under babble noise scenario.

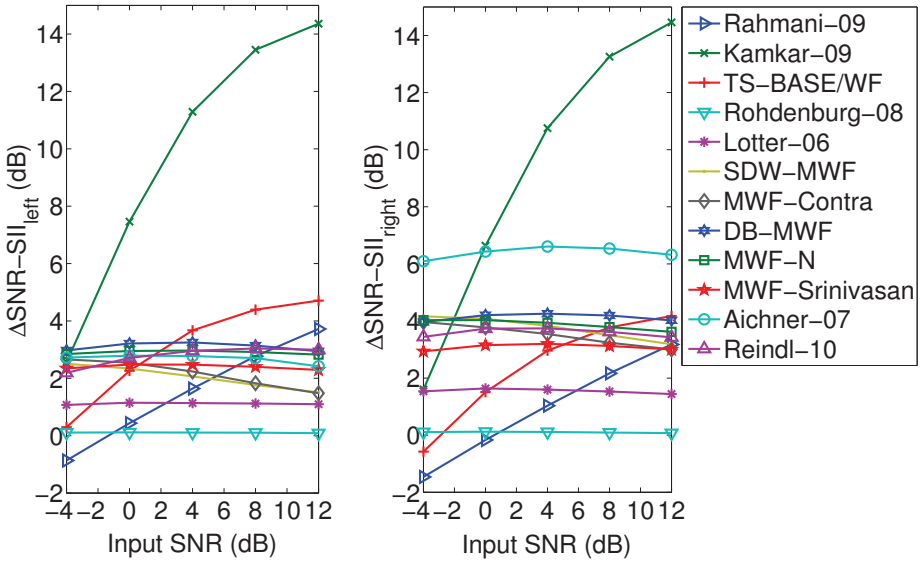

Figure 3: SNR improvement for existing techniques under multi-talker scenario.

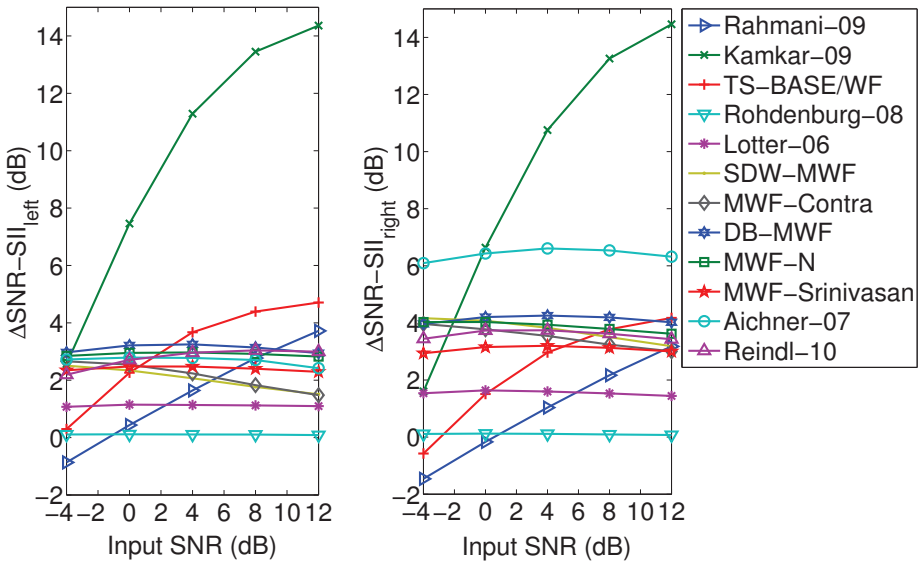

Figure 4: SNR improvement for existing techniques under moving source in babble noise scenario.

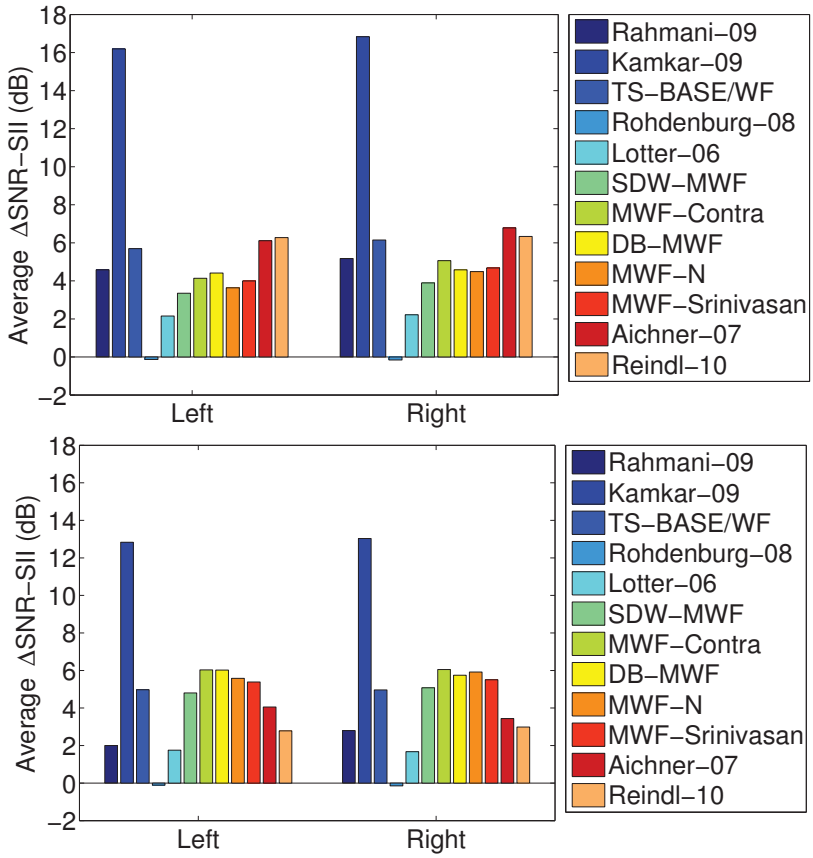

Figure 5: SNR improvement for existing techniques under variant-SNR babble noise: Getting in a cafeteria (Top) and Getting out a cafeteria (Bottom).

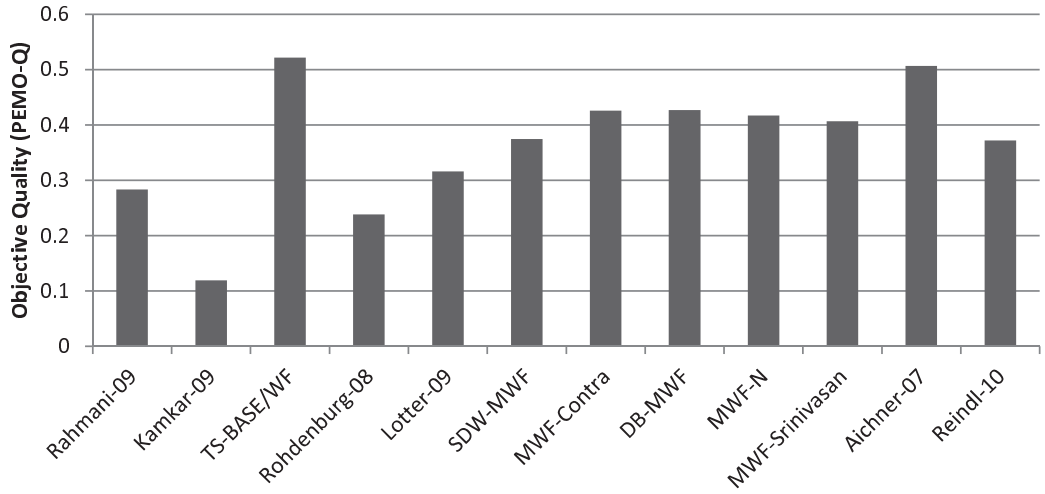

Figure 6: PEMO-Q scores for existing techniques under babble noise scenario at 0 dB input SNR. The highest PEMO-Q score is one, corresponding to a clean signal.

- Although TS-BASE/WF provides the best performance in terms of sound quality, its performance with respect to the SNR improvement is similar or lower than other techniques, e.g., MWF techniques. Particularly, for babble noise and multi-talker scenarios, its performance for low input SNR is very poor compared to other techniques.

- Beamforming algorithms (Rohdenburg-08 and Lotter-06) show the lowest performance among all techniques with respect to the SNR improvement (1-2dB).

- Taking into account the performance under all scenarios, MWF techniques provide the most desirable properties for a noise-reduction algorithm: Good output sound quality and high SNR improvement (3-7dB) at very low input SNR.

- All MWF techniques show similar performance. A theoretical analysis assuming stationary background noise shows that the SNR improvement in SDW-MWF should be higher than MWF-N [5]. However, our simulations show that this result is valid only under diffusive noise scenario. For other noise sources and scenarios, MWF-N and SDW-MWF provide similar performance. On the other hand, DB-MWF is shown to converge theoretically to SDW-MWF [13]; however our simulations show that DB-MWF outperforms SDW-MWF under all scenarios analyzed.

- Both BSS techniques considered in this study, Aichner-07 and Reindl-10, provide a

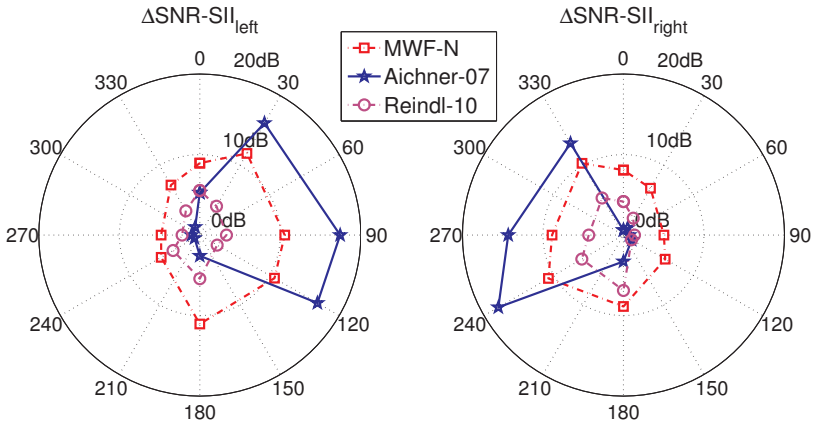

Figure 7: SNR improvement at different direction of arrivals of the target signal under babble noise scenario.

performance similar to the MWF techniques except under moving source scenario, and an acceptable sound quality. Although Aichner-07 outperforms Reindl-10 for all scenarios analyzed, we showed in [63] that the Aichner-07 method cannot preserve the localization cues for the interfering signals. In addition, an analysis regarding the performance of the SNR improvement at several direction of arrivals of the target signal shows that Aichner-07 is not a convenient approach. Figure 7 is a plot of the SNR improvement at different direction of arrival of the target signal under babble noise at 0 dB. The SNR improvement for other scenarios and SNR conditions exhibits similar shape. From this figure, it is clear that the ΔSNR-SII for both ears is almost symmetric in MWF-N and Reindl-10, but asymmetric in Aichner-07. This asymmetry in Aichner-07 is not a desirable feature in a binaural noise-reduction system because the perceptual quality of the sound is degraded. For example, if the direction of arrival of the target signal is 90^o, i.e., arriving at the right ear, Aichner-07 exhibits high noise reduction at the left ear but low noise reduction at the right ear. From a perceptual perspective, this means that the noise is not longer heard at the left ear, but it is still present at the right ear, producing an uncomfortable perception. This kind of behavior is not present in MWF-N and Reindl-10 because these techniques ensure similar noise reduction at both ears.

From the previous analysis, MWF and BSS techniques are shown to be the most convenient way to implement noise reduction in a binaural hearing aid because they provide acceptable SNR improvement at low and high SNR conditions, and good sound quality. Among all these techniques, MWF-N and Reindl-10 have the additional advantage that they preserve the localization cues for both target and interfering signals simultaneously, while the other MWF and BSS techniques only preserve the localization cues for the target signal. Moreover, MWF-N and Reindl-10 provide SNR improvement independent on the direction of arrival of the target signal.

DB-MWF is a method designed to reduce the transmission bandwidth. Results show that this method outperforms the other MWF methods. Hence, DB-MWF is a practical solution for those applications where the preservation of localization cues for the interfering signals is not important.

Since psycho-acoustic studies show the importance of preserving both localization cues, binaural noise-reduction methods such as DB-MWF are not convenient while other methods such as MWF-N and Reindl-10 are desirable. For this reason, this research proposed different methods based on BSS and MWF to accomplish the goals of performance improvement, computational-complexity simplification, and transmission-bandwidth reduction. The proposed methods are inspired in an auditory model. Taking into account perceptual properties is possible to achieve performance improvement while computational complexity and transmission bandwidth are reduced by removing unnecessary information from the perceptual viewpoint.

Two approaches are analyzed and discussed in the Chapters 4 and 5. The first approach is a BSS-based binaural noise-reduction method that uses an auditory filterbank to analyze the outputs of a BSS algorithm. Then, a set of time-domain gains are computed to expand the dynamic range of the input signals in a way similar to the mechanism used by the auditory system to adapt itself to a noisy environment (Chapter 4). The second approach is based on MWF, in which the FFT-based processing is replaced by an auditory representation (Chapter 5). The proposed MWF approach improves the performance of an FFT-based MWF and reduces computational complexity and transmission bandwidth.

Chapter IV

PERCEPTUALLY-INSPIRED BINAURAL NOISE REDUCTION USING BLIND SOURCE SEPARATION

The previous chapter showed that BSS-based and MWF-based noise-reduction methods are promising for binaural noise reduction. In these methods, MWF-N and Reindl-10 are the only two methods to preserve localization cues for both target and interfering signals and provide SNR improvement independent on the direction of arrival. However, a practical implementation of MWF-N and Reindl-10 involves block processing with large frame length, demanding long latency. For a hearing aid, latency is a critical parameter. Therefore, alternative methods to reduce the latency and to maintain or improve the performance are required. The first approach proposed in this research is based on blind source separation and perceptual post processing (BSS-PP). In the perceptual post processing, the input signals are analyzed in sub-bands and noise-reduction gains are computed for each sub-band, where the analysis fitlerbank and the expressions used to compute gains are inspired by an auditory model.

4.1 Background

When two-output BSS algorithms are used in noise-reduction applications, the primary BSS output provides an estimate of the target signal, and the secondary BSS output,

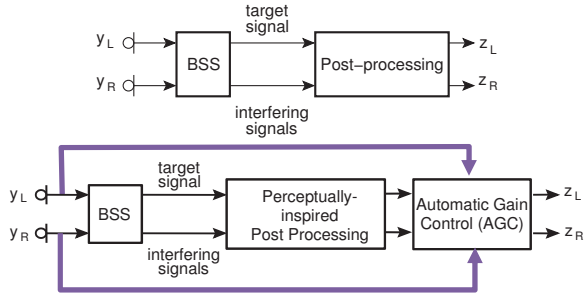

Figure 8: Top: Post processing in the existing BSS-based binaural noise-reduction methods. Bottom: Proposed post processing.

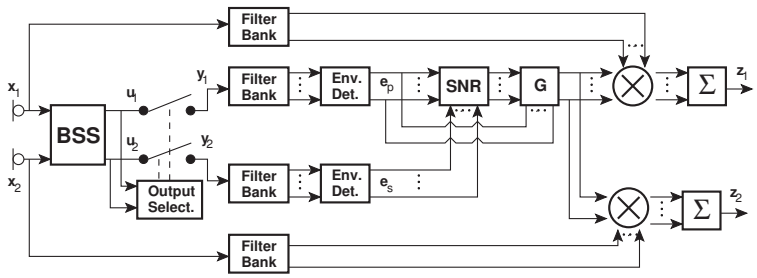

Figure 9: Proposed method based on BSS and perceptual post-processing.

an estimate of the interfering signal. In the existing BSS-based binaural noise-reduction methods, e.g., Aichner-07 [1], a post-processing stage is used to enhance the primary BSS output and recover the localization cues (Fig. 8a). In this research, a perceptually-inspired post processing is used to compute a set of time-domain gains from the BSS outputs, and these gains are applied to the unprocessed signals (Fig. 8b). The BSS post processing used in this work is an adaptation of the method in [72]. We selected this post processing since it outperforms other BSS post processing for monaural speech enhancement applications. This post processing is modified so that it can be used for a binaural hearing aid [62, 63]:

1. To preserve the localization cues, the gains obtained by the BSS and perceptual post-processing algorithm described in [72] are applied to the unprocessed signals received at each side (Figure 9).

2. To achieve low latency, the system is implemented assuming real-time operating constraints, with the envelopes (e_p and e_s), SNR estimates, and gain parameters updated in the frame-by-frame basis, while the gains and outputs are computed in the sample-by-sample basis. In [72], gains are computed assuming an entire knowledge of the signal.

3. To minimize artifacts and to achieve more quality outputs, it is necessary to hold a long-term history for the maximum values of the primary envelope (e_p). Different tests show that the length of this memory should be at least one second.

4. To estimate the SNR, first-order estimators of the signal and noise PSD are used, and the SNR is computed as the ratio of these PSDs.

4.2 Proposed Method (BSS-PP)

The proposed method is shown in Figure 9. Signals received at the left, x_1, and right, x_2, microphones are passed through a BSS algorithm to get u_1 and u_2. An output selection algorithm identifies which BSS output contains the separated target signal (y_1), or primary channel, and the separated interfering signal (y_2) or secondary channel. These outputs, y_1 and y_2, are analyzed using a constant-Q filterbank, and then, the envelope in each sub-band is extracted. These envelopes are used to estimate the SNR and to compute the noise-suppression gains. The SNR and gains are computed separately for each sub-band. These noise-suppression gains expand the dynamic range of each sub-band by lowering the noise floor. These gains are finally applied simultaneously to the unprocessed signals by time-domain multiplication, and the output from each sub-band is summed together to produce the enhanced signals for the left and right ear.

To reduce computational complexity and latency in the BSS stage, we use an info-max BSS algorithm that uses adaptive filters to minimize the mutual information of the system outputs. This algorithm is described by the following set of equations [63]:

$$u_1(n+1) = x_1(n) + \boldsymbol{w}_{12}^T(n)\boldsymbol{u}_2(n) \quad (2)$$

$$u_2(n+1) = x_2(n) + \boldsymbol{w}_{21}^T(n)\boldsymbol{u}_1(n) \quad (3)$$

$$\boldsymbol{w}_{12}(n+1) = \boldsymbol{w}_{12}(n) - 2\mu \tanh(u_1(n+1))\boldsymbol{u}_2(n) \quad (4)$$

$$\boldsymbol{w}_{21}(n+1) = \boldsymbol{w}_{21}(n) - 2\mu \tanh(u_2(n+1))\boldsymbol{u}_1(n)\,, \quad (5)$$

where x_1 and x_2 are the signals received at the left and right microphones, $\boldsymbol{w}_{12}$ and $\boldsymbol{w}_{21}$ are vectors of length N_w describing the unmixing filter coefficients, and $\boldsymbol{u}_1(n)$ and $\boldsymbol{u}_2$ are vectors of length N_w whose elements are the previous outputs of the BSS algorithm, $\boldsymbol{u}_j(n) = [u_j(n)\, u_j(n-1) \cdots u_j(n-N_w+1)]^T$, $j = 1, 2$, and n is the time index. To determine which BSS output contains the target signal, the time-average energy of the envelopes of the signals u_1 and u_2 are compared, and then, the output with higher time-average energy is selected as primary channel y_1. This time-average energy is computed by

$$u_j^{env}(n) = \eta_{env} u_j^{env}(n-1) + (1-\eta_{env}) u_j^2(n) \quad (6)$$

where η_{env} is a time constant. This update takes place every N samples.

The outputs of the BSS algorithm, $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$, as well as the unprocessed input signals at the left and right microphones, $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, are passed through a filterbank that resembles the auditory system. This filterbank was implemented using forth-order Butterworth filters. At 22 kHz sampling rate, each filterbank provides 24 sub-bands. At the output of the filterbanks, the vectors $\boldsymbol{x}_j(l,k)$ and $\boldsymbol{y}_j(l,k)$ of length N, $j = 1, 2$, are obtained, where l corresponds to the frame index and k to the sub-band number. Although the signals x and y are obtained in the sample-by-sample basis, they are analyzed in non-overlapped frames of length N in order to compute the gain parameters as we will show next.

For each output $\boldsymbol{y}_j(l,k)$, the envelope is extracted using a full-wave rectifier followed by a low-pass filter. In particular, the primary envelope vector $\boldsymbol{e}_p(l,k)$ is extracted from $\boldsymbol{y}_1(l,k)$, and the secondary envelope vector $\boldsymbol{e}_s(l,k)$ from $\boldsymbol{y}_2(l,k)$. The low-pass filters are implemented using a first-order IIR filter whose cutoff frequency is selected to be a fraction of the corresponding bandwidth of the band [72]. These cutoff frequencies are set to 1/5, 1/8 and 1/15 of the bandwidth of low, medium and high-frequency bands, respectively. These fractions ensure that the envelope tracks the signal closely but at the same time does not change too rapidly to cause abrupt gain changes that introduce modulation.

The final outputs at the left, z_1, and the right, z_2, side are computed using the time-domain gains $\boldsymbol{g}_{l,k}$ produced by the perceptual post-processing stage:

$$\boldsymbol{z}_j(l) \;=\; \sum_k \boldsymbol{g}_{l,k} \circ \boldsymbol{x}_j(l,k) \tag{7}$$

where $\circ$ denotes the element-wise product. The vector form emphasizes that the gains are computed using parameters updated on a frame-by-frame basis. However, these outputs can be computed on a sample-by-sample, reducing the latency.

In [72], inspired by a perceptual modeling, these gains modify the envelope of each sub-band $e_k(t)$ such that $\hat{e}_k(t) = \beta e_k^{\alpha}(t)$. To provide noise reduction, the maximum envelope value is preserved (i.e., $\hat{e}_{k_{max}} = e_{k_{max}}$) while the minimum envelope value is lowered (i.e., $\hat{e}_{k_{min}} = K e_{k_{min}}$, where K is an expansion coefficient). Using the previous ideas, [72] developed a method to estimate α and β from the entire signal. To provide a realistic

implementation, we modify the equations in [72] to a vector form to state the update of α and β is the frame-by-frame basis every N samples [62, 63]:

$$\boldsymbol{g}_{k,l} = \beta_{l,k} \boldsymbol{e}_p(l,k)^{(\alpha_{l,k}-1)}. \tag{8}$$

The factors α and β are computed as

$$\beta_{l,k} = \max(\boldsymbol{e}_{pmax}(k))^{(1-\alpha_{k,l})} \tag{9}$$

$$\alpha_{k,l} = 1 - \log K / \log M_{l,k}\,, \tag{10}$$

where $M_{l,k}$ is the SNR at k-th sub-band and l-th frame, and $\boldsymbol{e}_{pmax}(k)$, a vector that holds the maximum values of the primary envelopes, is obtained from the previous N_{max} frames:

$$\boldsymbol{e}_{pmax}(k) = [\max(\boldsymbol{e}_p(l,k)) \,...\, \max(\boldsymbol{e}_p(l-N_{max},k))] \tag{11}$$

To avoid computational overflow and preserve the binaural cues, the value of α is constrained in the range $\alpha = [0, 5]$. To minimize artifacts and achieve better quality outputs, the history stored in the vector $\boldsymbol{e}_{pmax}$ should hold at least one second. All experiments use two seconds of memory, i.e. $N_{max} = \lceil 2f_s/N \rceil$. Since α and β are fixed for a given frame, these gains can also be computed in the sample-by-sample basis.

To estimate the SNR at the given sub-band and frame, the signal and noise power are obtained from the envelopes of the primary and secondary channel. This approach reduces miss-classification errors in the SNR estimation when the input SNR is low. To obtain a reliable noise estimate, the noise power is updated using a rule derived from the noise PSD estimator proposed in [77]:

$$\begin{aligned}
&P_e = \|e_s(l,k)\|^2 \\
&if\ |P_e - P_v(l-1,k)| < \epsilon\sqrt{\sigma_v(l-1,k)} \\
&\quad P_v(l,k) = \lambda_v P_v(l-1,k) + (1-\lambda_v)P_e \\
&\quad \sigma_v(l,k) = \delta\sigma_v(l-1,k) + (1-\delta)\,|P_e - P_v(l-1,k)|^2 \\
&else
\end{aligned} \tag{12}$$

$$P_v(l,k) = P_v(l-1,k)$$

$$\sigma_v(l,k) = \sigma_v(l-1,k)$$

end

where $P_v(l,k)$ is the noise power at the k-th sub-band and l-th frame, $\sigma_v(l,k)$ is an estimate of the variance of P_v, λ and δ are time constants to smooth the estimation, and ϵ is a threshold coefficient. Finally, the frame SNR is estimated by

$$M_{l,k} = \max\left(\frac{P_x(l,k)}{P_v(l,k)} - 1\,,\, 1\right) \tag{13}$$

where P_x is the power of the primary channel estimated by

$$P_x(l,k) = \lambda_x P_x(l-1,k) + (1-\lambda_x)\left\|e_p(l,k)\right\|^2 \tag{14}$$

The values $\lambda_v = 0.95$, $\lambda_x = 0.9$, $\delta = 0.9$, and $\epsilon = 5$ provide good performance in our experiments.

The performance of the proposed algorithm depends on the tuning of two parameters: K and N. Whereas K controls the expansion of the dynamic range, N defines how often the parameters to compute the noise-suppression gains are updated. We presented a detailed analysis of the effect of these parameters on the SNR improvement and sound quality in [63]. In summary, $K = 0.01$ and $N = 8192$ are suitable for all scenarios [63].

4.3 Advantages and Limitations

In the proposed method, the noise-suppression gains are computed to expand the dynamic range of the noisy signal, in such a way that the maximum signal level is maintained while the noise level is pushed down. The maximum signal level is estimated from the primary channel, and the noise level from the secondary channel. Theoretical analysis conducted in [96] show that ICA-based BSS algorithms such as the algorithm used in our method provide an accurate noise estimate under non-point-source noise scenarios (e.g., diffusive or babble noise). Therefore, the performance of the proposed method under these scenarios is expected to be high. Since the proposed algorithm tracks the envelopes of the target speech and noise level simultaneously, it is expected to performance well under highly non-stationary environments. On the other hand, when the interfering signals are few point

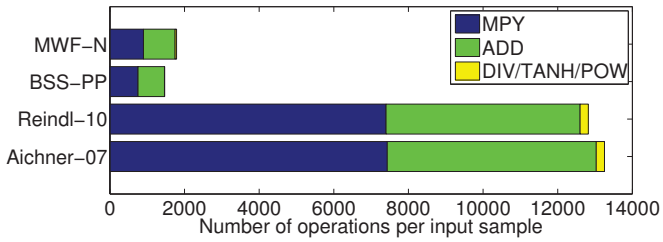

Figure 10: Number of operations for BSS-PP, MWF-N, Reindl-10, and Aichner-07 methods per input sample grouped into additions (ADD), multiplications (MPY), divisions (DIV), hyperbolic tangent (TANH), and power raise (POW).

sources, the BSS algorithm can provide accurate noise estimation only if the target signal is dominant. Thus, the performance of the proposed algorithm is expected to be low under these scenarios at very low input SNR. Fortunately, these kind of scenarios are uncommon. All the above statements are verified through experiments discussed in the Section 6.2. In general, the proposed method shows to be efficient to remove the background noise, provides an acceptable speech quality, preserves the localization cues for both target and interfering signals, and outperforms existing BSS-based and MWF-based methods (Section 6.2) in terms of SNR improvement and noise reduction.

Since the gains and outputs are computed in the sample-by-sample basis, the latency is very small (< 1 ms) compared to that of the existing methods (around 6 ms). In addition, the computational complexity of BSS-PP is slightly below the complexity of MWF-N and significantly smaller than Aichner-07 and Reindl-10 (Figure 10).

There are two main limitations in the proposed method. First, the subjective sound quality is acceptable, with a sound quality poorer than that of the MWF-N method. This suggests that MWF noise-reduction methods are more promising to achieve high noise reduction while maintaining the sound quality. This issue is addressed in the Chapter 5. Second, the BSS algorithm demands wireless transmission at full rate. Therefore, it is necessary to explore different reduced-bandwidth BSS algorithms, or to employ strategies other than BSS to estimate the target and interfering signals. This research adopted the second approach. In this case, a MWF-based framework is proposed to reduce the transmission bandwidth more aggressively than other existing MWF methods (Section 6.3.1).

Chapter V

PERCEPTUALLY-INSPIRED BINAURAL NOISE REDUCTION USING MULTICHANNEL WIENER FILTER

The comparative study conducted in this research (Chapter 3) showed that binaural noise-reduction methods based on BSS and MWF are promising techniques because of their performance under stationary and non-stationary environments. However, the existing BSS-based and MWF-based methods involve large latency and moderate computational complexity. To reduce latency and computational complexity, the previous chapter discussed a BSS-based binaural noise-reduction method. Although the proposed BSS-based method provides good noise reduction, the sound quality is acceptable, and the computational complexity and transmission bandwidth are not reduced. In this chapter, a MWF-based approach is discussed. This approach is based on perceptual information and provides good noise reduction and sound quality, and significant reduction in the computational complexity and transmission bandwidth. The method is called perceptual MWF (PMWF). In addition, this chapter discusses implementation strategies for reduction of the transmission bandwidth, improvement under non-stationary environments, and improvement of the speech intelligibility.

5.1 Background

In the FFT domain, for a particular frequency bin, f, and frame index, l, the signals received by all microphones can be described by the vector $\boldsymbol{y}(f,l) = [y_1^*(f,l)\, y_2^*(f,l)...y_{2M}^*(f,l)]^H$, where M is the number of microphones for each hearing aid. Assuming an additive background noise, the vector $\boldsymbol{y}(f,l)$ can be expressed as $\boldsymbol{y}(f,l) = \boldsymbol{x}(f,l) + \boldsymbol{v}(f,l)$, where $\boldsymbol{x}(f,l)$ and $\boldsymbol{v}(f,l)$ are the vectors that describe the target signal and noise components.

In the MWF framework, the filter coefficients are computed by minimization of the minimum mean square error (MMSE) between the filter outputs

$$z_L(f,l) = \boldsymbol{w}_L^H(f,l)\boldsymbol{y}(f,l) \quad (15)$$

$$z_R(f,l) = \boldsymbol{w}_R^H(f,l)\boldsymbol{y}(f,l) \quad (16)$$

and the desired signals, $x_L(f,l)$ and $x_R(f,l)$, where $\boldsymbol{w}_L$ and $\boldsymbol{w}_R$ are vectors of length $2M$ holding the filter weights, and the subscripts L and R represent the reference microphones at the left and right side. There are different MWF objective functions proposed in the literature [40, 13, 14] to obtain the filter weights. In particular, for the most widely-known MWF method, called speech distortion weighted MWF (SDW-MWF) [14], the weights at the right and left channel are computed by minimization of [14]

$$J_{SDW}(\boldsymbol{w}_L,\boldsymbol{w}_R) = \mathcal{E}\left\{\left\|\begin{bmatrix}(\boldsymbol{e}_L-\boldsymbol{w}_L)^H\boldsymbol{x}\\(\boldsymbol{e}_R-\boldsymbol{w}_R)^H\boldsymbol{x}\end{bmatrix}\right\|^2 + \mu\left\|\begin{bmatrix}\boldsymbol{w}_L^H\boldsymbol{v}\\\boldsymbol{w}_R^H\boldsymbol{v}\end{bmatrix}\right\|^2\right\}, \quad (17)$$

where μ denotes a trade-off parameter between noise reduction and speech distortion; $\boldsymbol{e}_L$ and $\boldsymbol{e}_R$ are unitary vectors of length $2M$ describing the position of the reference microphones for the left and right hearing aid. The indices f and l are dropped from the above equation for mathematical convenience. After minimization, the filter coefficients are given by [14]

$$\boldsymbol{w}_L(f,l) = (\boldsymbol{R}_x(f,l)+\mu\boldsymbol{R}_v(f,l))^{-1}\boldsymbol{R}_x(f,l)\boldsymbol{e}_L \quad (18)$$

$$\boldsymbol{w}_R(f,l) = (\boldsymbol{R}_x(f,l)+\mu\boldsymbol{R}_v(f,l))^{-1}\boldsymbol{R}_x(f,l)\boldsymbol{e}_R\,, \quad (19)$$

where $\boldsymbol{R}_x$ and $\boldsymbol{R}_v$ are the second-order statistics describing the speech and noise correlation matrices, defined as

$$\boldsymbol{R}_x(f,l) \triangleq \mathcal{E}\left\{\boldsymbol{x}(f,l)\boldsymbol{x}(f,l)^H\right\}$$

$$\boldsymbol{R}_v(f,l) \triangleq \mathcal{E}\left\{\boldsymbol{v}(f,l)\boldsymbol{v}(f,l)^H\right\}.$$

For practical implementations, the correlation matrix $\boldsymbol{R}_v$ can be estimated during the unvoiced segments using a voice activity detector (VAD). Under the assumption of statistical independence of the target and noise signals, the correlation matrix $\boldsymbol{R}_x$ can be estimated as $\boldsymbol{R}_x = \boldsymbol{R}_y - \boldsymbol{R}_v$, where $\boldsymbol{R}_y$ is estimated during voiced (or speech) segments. For practical

implementations, the matrices $\boldsymbol{R}_v$ and $\boldsymbol{R}_y$ can be updated using a first-order estimator

$$\boldsymbol{R}(f,l) = \alpha \boldsymbol{R}(f,l-1) + (1-\alpha)\boldsymbol{y}(f,l)\boldsymbol{y}^H(f,l)\,, \tag{20}$$

where $\boldsymbol{R}(f,l)$ is the correlation matrix (speech or noise), and α is a forgetting factor.

For MWF-N, the cost function includes an additional term to preserve a portion of the background noise [40]:

$$J_{MWFN}(\boldsymbol{w}_L, \boldsymbol{w}_R) = \mathcal{E}\left\{ \left\| \begin{bmatrix} (\boldsymbol{e}_L - \boldsymbol{w}_L)^H \boldsymbol{x} \\ (\boldsymbol{e}_R - \boldsymbol{w}_R)^H \boldsymbol{x} \end{bmatrix} \right\|^2 + \mu \left\| \begin{bmatrix} \eta \boldsymbol{e}_L^H \boldsymbol{v} - \boldsymbol{w}_L^H \boldsymbol{v} \\ \eta \boldsymbol{e}_R^H \boldsymbol{v} - \boldsymbol{w}_R^H \boldsymbol{v} \end{bmatrix} \right\|^2 \right\}. \tag{21}$$

After minimization, the equations to compute the filter weights in MWF-N are similar to (18) and (19),

$$\boldsymbol{w}_L(f,l) = (\boldsymbol{R}_x(f,l) + \mu \boldsymbol{R}_v(f,l))^{-1} \boldsymbol{R}_\eta(f,l)\boldsymbol{e}_L \tag{22}$$

$$\boldsymbol{w}_R(f,l) = (\boldsymbol{R}_x(f,l) + \mu \boldsymbol{R}_v(f,l))^{-1} \boldsymbol{R}_\eta(f,l)\boldsymbol{e}_R\,, \tag{23}$$

where

$$\boldsymbol{R}_\eta(f,l) = \boldsymbol{R}_x(f,l) + \mu\eta \boldsymbol{R}_v(f,l)\,, \tag{24}$$

and η is another trade-off parameter to control the amount of noise to be added at the output. This parameter takes the value $\eta = 0$ for SDW-MWF.

Practical real-time implementations of the MWF methods involve some challenges:

1. The implementation of FFT-based MWF involves high computational resources since the signal and noise correlation matrices as well as the weights are estimated for each frequency bin. For example, in SDW-MWF [14], having M microphones per hearing aid and using an FFT of length L, it is necessary to estimate $L/2$ correlation matrices of size $2M \times 2M$ for the signal and noise and to solve the same number of linear systems of equations.

2. The latency introduced by an FFT-based MWF processing depends on L. Since this delay is crucial for a digital hearing aid, it is necessary to keep L as small as possible.

Although both computational resources and latency may be reduced by smaller FFT lengths, Doclo *et al.* [13] have showed that the performance of the MWF methods is degraded when L is decreased.

3. Existing MWF methods require data transmission at full rate. Different methods to reduce the bandwidth are discussed in [13, 94]. Doclo *et al.* [13] discussed 4 approaches[1] that use a linear transformation to produce a single-channel signal to be transmitted over the link. Although the number of channels is reduced, the transmission is still at full rate. On the contrary, the method described in [94] splits the input signal into low-frequency and high-frequency components, and only the low-frequency component is transmitted. Then, each hearing aid uses monaural Wiener filters at high frequencies, and binaural Wiener filters at low frequencies.

4. The computation of the filter weights is very sensitive to second-order statistics estimation errors. A VAD-based estimation is not robust under highly-noisy and non-stationary environments [47] because VAD is very inaccurate for highly-noisy environments, and a VAD-based estimation is unable to track efficiently the statistics of non-stationary environments during voiced segments.

5. Existing noise-reduction methods cannot improve speech intelligibility for some low-input-SNR scenarios [46]. In [46], Loizou and Kim also showed that applying an ideal binary mask to the filter weights of a speech enhancement algorithm (spectral subtraction or Wiener filter) improves the speech intelligibility. However, the use of this binary mask degrades the sound quality.

To address these implementation challenges, this chapter presents different solutions:

- To reduce the computational complexity and latency, the FFT is replaced by an auditory representation. This auditory representation provides additional advantages such as better SNR improvement than the FFT-based processing, and the incorporation of aggressive strategies to reduce the transmission bandwidth. The proposed method

[1]The methods discussed in [13] are called DB-MWF, MWF-Contra, MWF-Front, and MWF-Superd. Among these methods, DB-MWF, or distributive binaural MWF, showed to provide the best performance.

called perceptual MWF (PMWF) is discussed in the Section 5.2 as well as the PMWF strategy to reduce the transmission bandwidth.

- To estimate the second-order statistics, a non-VAD method based on a multichannel noise cross-PSD (CPSD) is discussed in the Section 5.3. This method is combined with an adaptive estimation of the trade-off parameter μ to improve the noise reduction and sound quality.

- To improve the noise reduction at low-input SNR, a MWF framework that incorporates the auditory masking thresholds is discussed in the Section 5.4. This framework is shown to provide benefits only if the amount of estimation errors is small. In average, this method is outperformed by the CPSD-based method to estimate the second-order statistics (Section 6.3.3), and the CPSD-based method uses less computational resources than the method based on auditory masking thresholds.

- To improve speech intelligibility at very low input SNR, this research explores the feasibility of using binary masks in the binaural noise-reduction algorithm based on perceptual MWF. The proposed method, discussed in the Section 5.5, is shown to enhance speech intelligibility while maintain the SNR improvement, noise reduction, and sound quality of the original PMWF method.

5.2 Auditory Filterbank for Analysis/Synthesis

To reduce computational resources and latency, this research proposes an auditory processing instead of an FFT processing. The proposed method is called perceptual MWF or PMWF. An FFT-based processing can be seen as processing using an uniform filterbank (Fig. 11). In this case, increasing L increases the number of filters, reduces the bandwidth of each filter, and then, improves the frequency resolution at both low and high frequencies. However, it is widely known that the frequency response of the human auditory system is more selective at low frequencies (Fig. 11), and this low-frequency range is more relevant for speech intelligibility improvement. Therefore, in an FFT-based processing, to reach the low-frequency resolution of the human auditory system, it is necessary to employ large L.

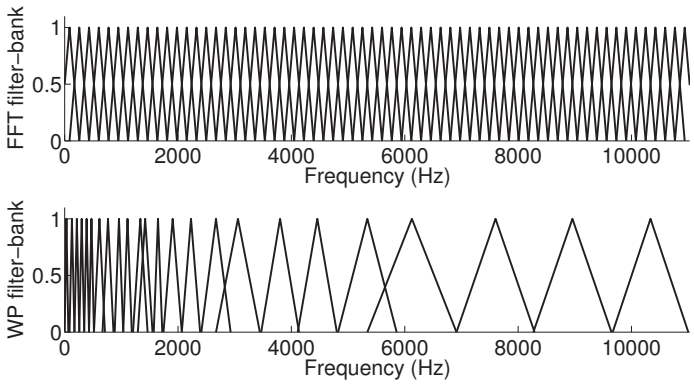

Figure 11: Filterbanks used in an FFT of length $L = 128$ and sampling frequency 22kHz (top) and a wavelet packet tailored to properties of the human auditory system (bottom).

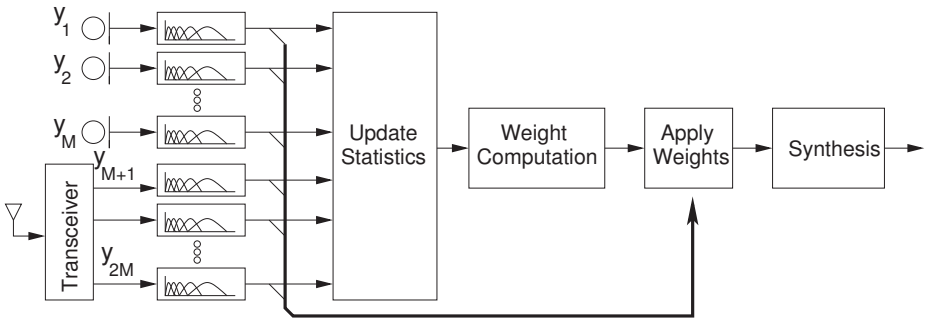

Figure 12: Proposed processing using auditory representation in MWF (PMWF).

Using larger L introduces more filters in the high-frequency range, but improving the SNR at these high-frequency bins does not contribute significantly to the speech intelligibility. Therefore, it is possible to reduce the computational cost without degrading the performance in a MWF method by using an auditory representation instead of an FFT representation. The additional advantage of this approach is that the number of sub-bands is fixed, and their respective bandwidths are independent of the frame length L. A typical number of auditory sub-bands is 20 for a sampling frequency of 16 kHz or 24 for $f_s = 22$ kHz. In the FFT-based MWF, a typically FFT length is $L = 128$, which meas 64 sub-bands. Therefore, the computational cost savings achieved with an auditory processing are very significant.

The proposed processing, in which the FFT has been replaced by an auditory filterbank is presented in Fig. 12. In this processing, the signals received at each microphone $y_i(n)$, $i = 1, ..., 2M$, are passed through an auditory filterbank, which decomposes each input signal

into K sub-bands. Each sub-band output is represented by $y_{i,k}(n_k)$, where k corresponds to the sub-band index, and n_k is a time-index in the auditory domain. If the operation applied by the auditory filterbank is linear, it is possible to describe $y_{i,k}$ as $y_{i,k} = x_{i,k} + v_{i,k}$, where x and v correspond to the target and noise components, respectively. In other words, the equations for the SDW-MWF framework, (18) and (19), are still valid, where (f, l) is replaced by (k, n_k). Three ways to implement the auditory-domain transformation are explored in this research: IIR filterbank (FB) [55], wavelet packet (WP) [54, 55], and frequency-warped filters (FW) [59, 58].

5.2.1 Implementation Based on IIR fitlerbank (FB-PMWF)

In the IIR filterbank (FB) implementation, constant-Q 4th-order IIR filters are used for analysis. This filterbank is designed to approximate a critical band specification [103]. Hence, for 22 kHz sampling rate, the filterbank provides 24 sub-bands, and for 16 kHz sampling rate, 20 sub-bands. The weights are computed for each sub-band and frame, and the synthesis is performed by adding the weighted filterbank outputs.

The FB-based PMWF has the advantage of providing very small latency. However, the statistics and weights are updated at full rate, which is computational expensive. To reduce complexity, the statistics are updated in the frame-by-frame basis. If L is the frame length, the statistics for the frame index, l, and sub-band, k, can be the updated by a first-order estimator [54]:

$$\boldsymbol{R}(k,l) = \alpha R(k,l-1) + \frac{1}{L_k}(1-\alpha)\boldsymbol{Y}(k,l)\boldsymbol{Y}(k,l)^T \tag{25}$$

where $\boldsymbol{Y}(k,l)$ is a $2M \times L_k$ matrix with the auditory representation at the k-th sub-band and frame index l; and L_k is the number of samples at this sub-band. For FB-PMWF, $L_k = L \ \forall k$. Using the correlation matrices defined in (25), the filter weights at the left and right channel, $\boldsymbol{w}_L$ and $\boldsymbol{w}_R$, are applied to all samples within the frame, i.e.,

$$\boldsymbol{z}_L(k,l) = \boldsymbol{w}_L^H(k,l)\boldsymbol{Y}(k,l) \tag{26}$$

$$\boldsymbol{z}_R(k,l) = \boldsymbol{w}_R^H(k,l)\boldsymbol{Y}(k,l) \tag{27}$$

where $\boldsymbol{z}_L(k,l)$ and $\boldsymbol{z}_R(k,l)$ are vectors of length L_k holding the output samples for the k-th sub-band and l-th frame.

FB-PMWF provides a performance similar or worse than WP-PMWF, and WP-PMWF uses less computational resources (Section 6.3) [55]. For this reason, this document focuses extensively on WP-PMWF.

5.2.2 Implementation Based on Wavelet Packet and Reduction of Transmission Bandwidth (WP-PMWF)

There are different WP trees proposed in the literature to imitate the human auditory system. This research uses the WP tree proposed in [32]. We have reported the proposed method for mother wavelet Daubechies 4 (db4) and 8 (db8) [54, 55], where both mother wavelets provide similar performance, but db4 involves less computational resources. A detailed analysis of the effect of different mother wavelet is addressed in the Section 6.3.

The WP-based implementation also ensures the absence of block-processing artifacts due to the perfect reconstruction inherent in the WP, which is not the case of the FB-based implementation.

As a result of the multirate processing inherent in the WP computation, the computational complexity and transmission bandwidth can be reduced significantly compared to the FFT and FB implementations. In the WP-based implementation, the number of samples at each sub-band, L_k, is a is a power-of-two fraction of the frame length L, which reduces the size of the matrices $\boldsymbol{Y}(k,l)$ in (25), and so the number of operations required to update $\boldsymbol{R}(k,l)$. Again, (18) and (19) are used to compute the filter weights, and (26) and (27) to compute the WP representation of the output. Moreover, an FFT-based implementation involves complex-valued operations while PMWF only real-valued operations.

Existing reduced-bandwidth MWF methods proposed by Doclo *et al.* [13] reduce the number of channels to be transmitted over the link to one channel, but the transmission of this single channel is required at full-rate. Transmission bandwidth may be reduced by using a suitable channel codification, but the codification itself may increase the computational complexity. The method proposed in this Section reduces both transmission bandwidth and computational cost (Section 6.4).

To preserve the localization cues, the interaural time differences (ITD) and the interaural level differences (ILD) should be preserved. Whereas ITD cues are more relevant for

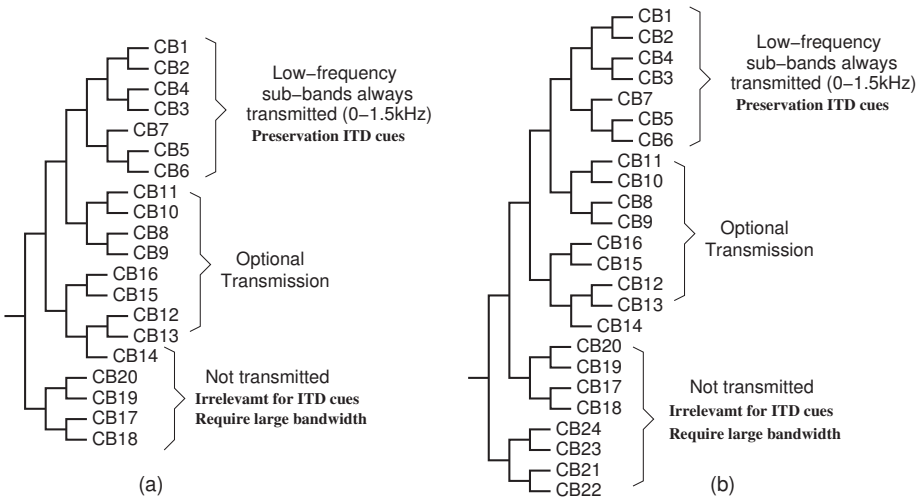

Figure 13: Proposed bandwidth reduction in PMWF. Wavelet tree for a sampling frequency of (a) 16 kHz and (b) 22 kHz.

frequencies below 1.5 kHz, ILD cues are more relevant for high frequencies. However, the ITD cues are more important than the ILD for the identification of the direction of arrival [65]. Hence, to preserve the ITD cues is necessary to preserve the phase information between the processing at the left and the right hearing aid, i.e., it is necessary to use binaural processing. These ideas were initially explored by Srinivasan [94], in which a binaural Wiener filter is used only for the low frequency region while a monaural Wiener filter is used for the high frequency region. In the proposed method, the outputs of the low-frequency sub-bands are signals at very slow rate. Thus, transmitting the information of the sub-bands related to frequencies below 1.5 kHz is an efficient way to reduce the transmission bandwidth in the PMWF method. These ideas are illustrated in the Fig. 13 for the wavelet tree employed in PMWF.

5.2.3 Implementation Based on Frequency-Warped Filters (FW-PMWF)

Frequency-warped DFT (WDFT) is another way to implement an auditory filterbank [71, 35]. In this context, the input signal is passed through a chain of first-order all-pass filters to obtain a set of signals having deformed spectral components. Then, a DFT is taken to this set of signals to obtain the warped spectrum (Fig. 14). Each all-pass filter has a transfer function given by

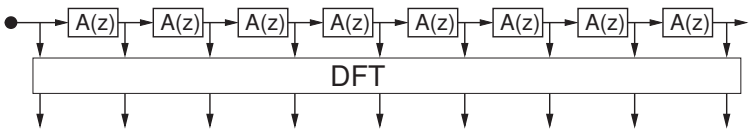

Figure 14: Warped Discrete-Fourier Transform (WDFT).

$$A(z) = \frac{z^{-1} - a}{1 - az^{-1}} \tag{28}$$

where a is the warping parameter. A suitable selection of this warping parameter provides a spectrum close to the auditory representation [71, 88]. The value of a depends on the sampling frequency [88]. For example, for a sampling frequency of 16 kHz, $a = 0.5756$. A WDFT-based filterbank as described in [50] can be used to implement the analysis and synthesis stages in the PMWF method. In this case, the correlation matrices, filter weights, and frequency-warped representation of the output can be computed with the same expressions used for the WP-based implementation. Similar to the FB-based implementation, the performance of this WDFT-based implementation is close to the performance of the WP-based implementation but the WDFT-based implementation uses more computational resources than the WP-based implementation. Hence, the WP-based implementation is still the most promising implementation for the PMWF method.

Kates and Arehart [35] proposed a frequency-warped FIR filter to implement a wide dynamic range compression (WDRC) system. In this case, the filter coefficients are computed in the frequency-warped domain, and applied in the time-warped domain, i.e., the output is a linear combination of the all-pass filter outputs. This structure has small latency, and the group delay is frequency dependent, with large group delay at low frequency and small group delay at high frequency.

In this research, frequency-warped FIR filters are proposed for the MWF implementation. In this case, $2M$ frequency-warped FIR filters are used to process the signals received at each microphone and combined to obtain the enhanced outputs. The coefficients of the frequency-warped filters are computed using the SDW-MWF framework. These coefficients can be computed using the information in the frequency-warped domain (or WDFT domain) or in the time-warped domain (i.e., from the all-pass filter outputs) (Fig. 15) by

expressions similar to (18) and (19). The output of the FW-PMWF method is given by

$$z_L(n) = \boldsymbol{w}_L^H(n)\, vec(\tilde{\boldsymbol{Y}}_n) \tag{29}$$

$$z_R(n) = \boldsymbol{w}_R^H(n)\, vec(\tilde{\boldsymbol{Y}}_n)\,, \tag{30}$$

where the matrix $\tilde{\boldsymbol{Y}}_n = \{\tilde{y}_{k,m}(n)\}$, and $\tilde{y}_{k,m}(n)$ is the output of the k-th all-pass filter in the chain of the m-th input microphone at the time index n. This matrix has a size $K \times 2M$. For the time-warped FW-PMWF method, weights are computed by

$$\boldsymbol{w}_L(n) = \left(\boldsymbol{R}_{\tilde{X}}(n) + \mu \boldsymbol{R}_{\tilde{V}}(n)\right)^{-1} \boldsymbol{R}_{\tilde{X}}(n)\boldsymbol{e}_L \tag{31}$$

$$\boldsymbol{w}_R(n) = \left(\boldsymbol{R}_{\tilde{X}}(n) + \mu \boldsymbol{R}_{\tilde{V}}(n)\right)^{-1} \boldsymbol{R}_{\tilde{X}}(n)\boldsymbol{e}_R\,, \tag{32}$$

where

$$\boldsymbol{R}_{\tilde{X}}(n) \triangleq \mathcal{E}\left\{vec(\tilde{\boldsymbol{X}}_n)vec(\tilde{\boldsymbol{X}}_n)^H\right\}$$

$$\boldsymbol{R}_{\tilde{V}}(n) \triangleq \mathcal{E}\left\{vec(\tilde{\boldsymbol{V}}_n)vec(\tilde{\boldsymbol{V}}_n)^H\right\}.$$

For the frequency-warped FW-PMWF method,

$$\boldsymbol{w}_L(n) = \left(\boldsymbol{I}_{2M} \otimes \boldsymbol{F}^{-1}\right)\left(\boldsymbol{R}_{\hat{X}}(n) + \mu \boldsymbol{R}_{\hat{V}}(n)\right)^{-1} \boldsymbol{R}_{\hat{X}}(n)\left(\boldsymbol{I}_{2M} \otimes \boldsymbol{F}\right)\boldsymbol{e}_L \tag{33}$$

$$\boldsymbol{w}_R(n) = \left(\boldsymbol{I}_{2M} \otimes \boldsymbol{F}^{-1}\right)\left(\boldsymbol{R}_{\hat{X}}(n) + \mu \boldsymbol{R}_{\hat{V}}(n)\right)^{-1} \boldsymbol{R}_{\hat{X}}(n)\left(\boldsymbol{I}_{2M} \otimes \boldsymbol{F}\right)\boldsymbol{e}_R\,, \tag{34}$$

where

$$\boldsymbol{R}_{\hat{X}}(n) \triangleq \mathcal{E}\left\{vec(\hat{\boldsymbol{X}}_n)vec(\hat{\boldsymbol{X}}_n)^H\right\}$$

$$\boldsymbol{R}_{\hat{V}}(n) \triangleq \mathcal{E}\left\{vec(\hat{\boldsymbol{V}}_n)vec(\hat{\boldsymbol{V}}_n)^H\right\},$$

and $\tilde{\boldsymbol{Y}}_n = \tilde{\boldsymbol{X}}_n + \tilde{\boldsymbol{V}}_n$, $\hat{\boldsymbol{Y}}_n = \boldsymbol{F}\tilde{\boldsymbol{Y}}_n$, and $\boldsymbol{F}$ denotes the Fourier transform operator, $vec(\boldsymbol{A})$ represents the vectorization of the matrix $\boldsymbol{A}$, and $\otimes$ is the tensor product. A complete derivation of these equations is presented in the Appendix A.

Results showed that the MWF framework derived in the time-warped domain provides better performance and less computational cost than the frequency-warped domain (Section

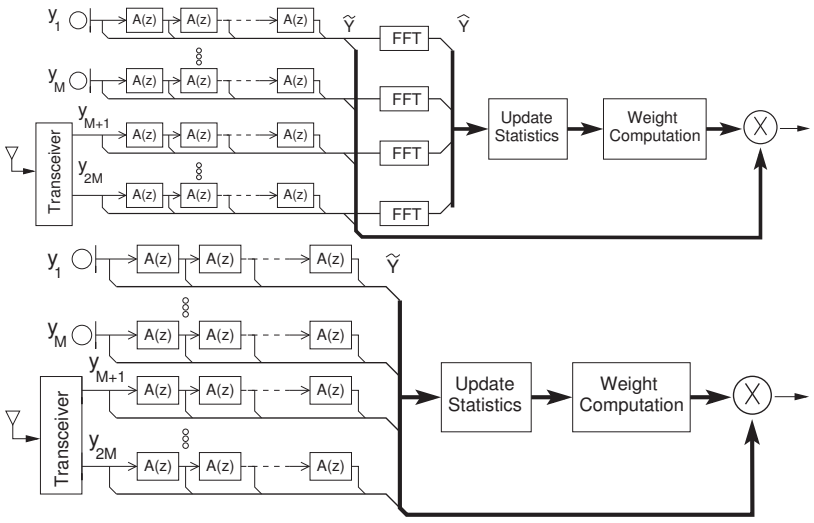

Figure 15: Frequency-warped MWF (FW-PMWF). Top: FW-PMWF coefficients derived in the frequency-warped domain. Bottom: FW-PMWF coefficients derived in the time-warped domain.

6.3.2). In addition, compared to the WP-based implementation, the FW-based implementation provides similar performance in terms of SNR improvement, noise reduction, and sound quality, but the FW-based implementation uses more computational resources than the WP-based implementation (Section 5.6).

5.3 Second-Order Statistics Estimation Based on Multichannel Noise Cross-PSD (MWF-CPSDμ_{SNR})

MWF approaches require the estimation of second-order statistics to compute the filter weights. These statistics have been typically estimated using a VAD-based method, in which the noise statistics are updated during noise-only segments, and signal statistics during voiced segments. VAD-based second-order statistics estimation is challenging for highly-noisy and non-stationary environments [47]. There are experimental [13] and theoretical [6] evidences that VAD errors degrade significantly the performance of MWF-based noise-reduction methods. Even using a perfect VAD, a VAD-based implementation is unable to track efficiently the statistics of non-stationary environments during voiced segments. These reasons motivate the exploration of non-VAD-based implementations for MWF such as the method proposed in this section, which we discussed in detail in [56].

In the proposed method, second-order statistics are estimated using a multichannel

noise cross-PSD (CPSD) estimator [56]. We analyzed three estimation methods in [56], concluding that a weighted spectral averaging method offers the best performance. The proposed multichannel weighted spectral averaging method is an extension of the method in [77]:

$$
\begin{aligned}
&\boldsymbol{R}_t(f) = \boldsymbol{y}(f,l)\boldsymbol{y}^H(f,l) \\
&\text{if} \quad tr(|\boldsymbol{R}_t(f) - \boldsymbol{R}_v(f,l-1)|) < \epsilon\sqrt{\sigma(f,l-1)} \\
&\quad \boldsymbol{R}_v(f,l) = \alpha_v \boldsymbol{R}_v(f,l-1) + (1-\alpha_v)\boldsymbol{R}_t(f) \\
&\quad \sigma(f,l) = \delta\sigma(f,l-1) + (1-\delta)tr(|\boldsymbol{R}_t(f) - \boldsymbol{R}_v(f,l-1)|)^2 \\
&\quad \boldsymbol{R}_x(f,l) = \alpha_x \boldsymbol{R}_x(f,l-1) \\
&\text{else} \\
&\quad \sigma(f,l) = \sigma(f,l-1) \\
&\quad \boldsymbol{R}_x(f,l) = \alpha_x \boldsymbol{R}_x(f,l-1) + (1-\alpha_x)\boldsymbol{R}_t(f) \\
&\text{end}
\end{aligned}
$$

This method differs from [77] in the use of the trace–$tr()$ to estimate the variance of the estimation; α and δ correspond to forgetting and smoothing factors, respectively. The decision rule to update $\boldsymbol{R}_v$ corresponds to the detection of a noise-only time-frequency bin. The main difference between the proposed method and a VAD-based estimation is the update of individual time-frequency bins rather than all time-frequency bins according to the frame class. Thus, the proposed method can track the noise dynamics during voiced segments for those frequency bins where the speech signal is being masked.

A better performance in SDW-MWF is obtained by adapting the trade-off parameter μ. This parameter controls the amount of noise reduction and speech distortion. Thus, using two values of μ, one for noise-only segments and another one for voiced segments, may improve the noise reduction and reduce the speech distortion. Based on this principle, Ngo *et al.* [67] proposed an adaptive μ that uses the inverse of the probability of being a voiced segment. This probability is determined by a soft-VAD. This method is shown to improve the robustness against VAD-induced errors. For single-channel Wiener filters, μ can be adapted according to frame SNR, auditory masking thresholds, or perceptual weighting [47]. Due to the simplicity of an adaptive μ based on frame SNR, this is an ideal solution

for a hearing aid. A typical way to estimate μ, based on SNR, is given by [47]

$$\mu = \mu_0 - SNR_{dB}/s\,, \tag{35}$$

where s is a scaling factor. This equation has some limitations for low-input SNR [56]. Particularly, we conducted in [56] that smaller μ must be employed at low-input SNR to avoid large speech distortion. For this reason, in [56], a method to adapt μ is derived for a multichannel MWF. For a single target signal, the speech correlation matrix can be expressed as [5]

$$\boldsymbol{R}_x(f,l) = P_s(f,l)\boldsymbol{a}_{f,l}\boldsymbol{a}_{f,l}^H\,, \tag{36}$$

where $P_s(f,l)$ is the target signal power, and $\boldsymbol{a}_{f,l}$ the propagation vector for the target signal. Replacing (36) in (18) and after some manipulations,

$$\boldsymbol{w}_L(f,l) = \frac{P_s(f,l)}{\xi(f,l)+\mu(f,l)}\boldsymbol{R}_v^{-1}(f,l)\boldsymbol{a}_{f,l}\boldsymbol{a}_{f,l}^H\boldsymbol{e}_L\,,$$

where $\xi(f,l) = P_s(f,l)\boldsymbol{a}_{f,l}^H\boldsymbol{R}_v^{-1}(f,l)\boldsymbol{a}_{f,l}$ corresponds to the output SNR at the frequency f and frame index l. In the above equation, we also assume that μ depends on the frequency and time. A similar equation is obtained for the right side. Applying a constraint to the output noise power

$$\begin{aligned}\boldsymbol{w}_L^H(f,l)\boldsymbol{R}_v(f,l)\boldsymbol{w}_L(f,l) &< P_{v,des}\\ \frac{\xi(f,l)}{(\xi(f,l)+\mu(f,l))^2}\left|\boldsymbol{a}_{f,l}^H\boldsymbol{e}_L\right|^2 &< P_{v,des}/P_s(f,l)\,,\end{aligned}$$

where $P_{v,des}$ is the desired output noise power. Since $\left|\boldsymbol{a}_{f,l}^H\boldsymbol{e}_L\right| \leq 1$, it is possible to establish a condition for μ based on the output SNR, ξ, and the desired output SNR, $\xi_{des} = P_s(f,l)/P_{v,des}\forall f$,

$$\mu(f,l) \geq \sqrt{\xi(f,l)\cdot\xi_{des}} - \xi(f,l)\,. \tag{37}$$

Hence, to minimize speech distortion, trade-off parameters for the left and right sides, μ_L and μ_R, can be adapted by

$$\begin{aligned}\mu_L(f,l) &= \sqrt{\xi_L(f,l)\xi_{des}} - \xi_L(f,l)\\ \mu_R(f,l) &= \sqrt{\xi_R(f,l)\xi_{des}} - \xi_R(f,l)\end{aligned} \tag{38}$$

The estimation of the output SNR, ξ, involves a large number of operations. However, different test showed that this term can be replaced by an estimate of the *a priori* SNR, which is computed for the left reference microphone as $\xi_L(f,l) = \boldsymbol{e}_L^H \boldsymbol{R}_x(f,l)\boldsymbol{e}_L / \boldsymbol{e}_L^H \boldsymbol{R}_v(f,l)\boldsymbol{e}_L$.

The above strategy is referred as MWF-CPSDμ_{SNR} in the remaining of this document.

5.4 MWF Framework Based on Auditory Masking Thresholds (MWF-μ_{ATH})

In monaural speech enhancement applications based on Wiener filter, auditory masking thresholds can be used to limit the output noise power level. In this sense, the noise power level is reduced to a value lower than the auditory masking threshold [47]. It is also known that the noise reduction in this monaural Wiener filter is controlled by a trade-off parameter depending on the auditory masking threshold. This trade-off parameter controls the amount of noise reduction and speech distortion in the same way as the parameter μ in SDW-MWF. To incorporate the auditory masking thresholds in the MWF framework, a cost function that minimizes the speech distortion while constrains the noise power to a level below the auditory masking threshold T_f is proposed as follows:

$$\min \mathcal{E} \left\{ \left\| \begin{array}{c} (\boldsymbol{e}_L - \boldsymbol{w}_L)^H \boldsymbol{x} \\ (\boldsymbol{e}_R - \boldsymbol{w}_R)^H \boldsymbol{x} \end{array} \right\|^2 \right\} \; subject\, to \left\| \begin{array}{c} \boldsymbol{w}_L^H \boldsymbol{v} \\ \boldsymbol{w}_R^H \boldsymbol{v} \end{array} \right\|^2 < T_f. \tag{39}$$

This framework leads to the same set of equations as SDW-MWF, (18) and (19), and the trade-off parameter μ is a non-linear function of the auditory masking threshold [47]:

$$\mu(f,l) = \max \left(\sqrt{\frac{P_x(f,l)}{T_f \xi(f,l)}} - \xi(f,l), 0 \right), \tag{40}$$

where $P_x(f,l)$ is the power of the clean signal at the frequency bin f and frame index l, T_f is the auditory masking threshold, and $\xi(f,l)$ is the *a priori* SNR. In practical implementations, the estimation of the noise level is more accurate compared to the estimation of the speech power (P_x). Therefore, the equation (40) may fail for highly-noisy environments. An alternative solution to (40) is to use an iterative method that checks the noise power level. If the noise power level is above the auditory masking threshold, μ must be

increased, otherwise μ must be decreased. The above strategy is referred as MWF-μ_{ATH} in the remaining of this document.

Equation (40) has a close relationship with the expression for the adaptive μ proposed for the MWF-CPSDμ_{SNR} framework (Section 5.3). Particularly, the expression proposed in Section 5.3, equation (38), replaces the ratio $P_x(f,l)/T_f$ by a desired SNR, ξ_{des}. Thus, MWF-CPSDμ_{SNR} involves less computational cost than MWF- μ_{ATH}, and it may avoid uncertainties in the estimation of both speech power and auditory masking threshold. These issues are further discussed in the Section 6.3.3, concluding that MWF-CPSDμ_{SNR} provides similar performance to MWF-μ_{ATH}. Hence, MWF-CPSDμ_{SNR} is sufficient for a real-time implementation of a MWF-based binaural noise-reduction method.

5.5 Improvement of Speech Intelligibility by Binary Masking (MWF-IDBM)

To improve speech intelligibility in noisy signals, some authors have shown that ideal binary masks (IDBM) applied in the time-frequency domain can improve the intelligibility significantly [2, 45]. The idea behind IDBM is that those time-frequency (T-F) bins where the masker signal (speech) is stronger than the noise signal must be preserved (i.e., when local SNR $>$ threshold), and those T-F bins where the noise is stronger, must be removed (i.e., when local SNR $<$ threshold). A recent study addressed the reasons why speech intelligibility is not improved by some monaural speech enhancement algorithms [46]. In [46], Loizou and Kim identified that the existing algorithms may provide solutions in a region where speech distortion leads to degradation in speech intelligibility. In addition, these authors showed that applying an ideal binary mask to the filter weights computed by the speech enhancement algorithm (spectral subtraction or Wiener filter) improves the speech intelligibility. Although, applying an IDBM to the unprocessed signal improves the speech intelligibility, it introduces musical-noise artifacts and distortions in the background noise. These artifacts and distortions are absent in MWF, but MWF does not provide significant speech intelligibility improvement as IDBM as will be shown in Section 6.3.6. Hence, this research studied the feasibility of using binary masks to improve speech intelligibility in the binaural noise-reduction algorithm described in Section 5.2.2.

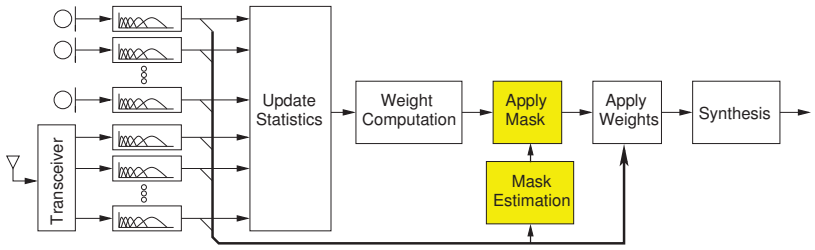

Figure 16: Block diagram of the proposed solution to improve speech intelligibility in MWF. The shaded blocks are the proposed modification to improve speech intelligibility.

The proposed method is the combination of a MWF and an IDBM algorithm, in which a post processing based on binary mask is introduced to the MWF weights (Figure 16). The intention of this combination is to obtain a technique with the advantages of MWF with respect to noise reduction and sound quality, and the advantages of IDBM with respect to speech intelligibility improvement. In the proposed method, called MWF-IDBM, the MWF weights to filter out the input signals are given by

$$\hat{\boldsymbol{w}}_L(f,l) = g_L(f,l)\boldsymbol{w}_L(f,l) \tag{41}$$

$$\hat{\boldsymbol{w}}_R(f,l) = g_R(f,l)\boldsymbol{w}_R(f,l) \tag{42}$$

where $\boldsymbol{w}_L(f,l)$ and $\boldsymbol{w}_R(f,l)$ are the weights for the left and right channel, respectively, computed by the MWF algorithm at the frequency bin f and frame index l. $g_L(f,l)$ and $g_R(f,l)$ are the binary masks for the left and right channel, respectively, and $\hat{\boldsymbol{w}}_L(k,l)$ and $\hat{\boldsymbol{w}}_R(k,l)$ are the weights to enhance the signal. The equations (41) and (42) can be derived from the ideas described in [46]. This derivation is included in the Appendix B. If the masks g_L and g_R take the values $\{\eta, 1\}$, with $\eta \approx 0$, the proposed method can be seen as an IDBM algorithm using soft mask, where the soft mask corresponds to the MWF weights. This soft mask avoids discontinuities in the output time-frequency spectrum, reducing the audible artifacts introduced by the IDBM algorithm.

The computation of the ideal binary mask is based on the local SNR, $\xi(f,l)$,

$$g(f,l) = \begin{cases} 1 & \xi(f,l) > \theta \\ \eta & otherwise \end{cases} \tag{43}$$

The value of the threshold θ is chosen to be 0 dB. This value has been found to be close

the boundary where the speech intelligibility starts decreasing [45]. η is a parameter to control the amount of audible artifacts. This parameter must be chosen close to zero to obtain speech intelligibility improvement. A value of $\eta = 0.1$ is found to be suitable for most scenarios. Binary masks proposed in the literature are proposed for a monaural case [4, 22, 80, 45, 100, 38]. For a binaural mask generation, different strategies are analyzed. In these strategies, independent ideal binary masks, $\hat{g}_L$ and $\hat{g}_R$, are estimated for each side, e.g., the left mask, $\hat{g}_L$, is estimated using the local SNR at the left channel. These strategies are as follow: a) use independent binary masks for each channel, i.e., $g_L = \hat{g_L}$ and $g_R = \hat{g}_R$; b) AND combination of the independent ideal masks, i.e., $g_L = g_R = \hat{g}_L \cdot \hat{g}_R$; c) OR combination of the independent ideal masks, i.e., $g_L = g_R = \hat{g}_L + \hat{g}_R$.

Objective metrics and informal listening tests showed that the mask generation strategy based on independent masks provide the best performance in terms of noise reduction, sound quality, and speech intelligibility (Section 6.3.6), and the AND-combined mask degrades the performance of the MWF component.

Under ideal conditions, i.e., ideal binary mask and perfect VAD for the estimation of the MWF statistics, the proposed method improves speech intelligibility under highly non-stationary environments for input SNR $\leq 0\,dB$, and it avoids the distortion artifacts present in a standalone IDBM method (Section 6.3.6). Different on-line strategies to generate the binaural binary mask are discussed in the Section 6.3.6. However, these strategies could not provide a performance near to the ideal case, which suggests further research in this field.

5.6 Advantages and Limitations

In this chapter, different MWF processing strategies were proposed. These strategies are based on the replacement of the FFT processing by an auditory filterbank. The proposed method, called PMWF, is implemented using an IIR filterbank (FB), wavelet packet (WP), or frequency-warped filters (FW). A summary of the features of the FFT-based processing and the proposed processing is presented in the Table 2. Among the PMWF methods, the WP-PMWF method exhibits preferable features for a practical implementation: low computational complexity, low latency, and availability of methods to reduce the transmission

bandwidth. In terms of performance, all implementations provide similar performance and outperform the FFT-based MWF implementation as will be shown in the next chapter. This fact is explained by the usage of an auditory filterbank that increases the frequency resolution at low frequency compared to the usage of an FFT-based processing.

This chapter also introduced strategies to update the second-order statistics and the trade-off parameter μ. With respect to the second-order statistics estimation, a multi-channel CPSD-based method is introduced to replace the well-known VAD-based method. Different from a VAD-based estimation method, the proposed method is designed to track changes in the noise statistics during voiced segments. Hence, the proposed method is expected to provide a good performance under highly non-stationary environments. This claim will be verified through different experiments in the next chapter (Section 6.3.3). With respect to the trade-off parameter μ, this research proposes two methods to adapt μ: using target SNR or auditory masking thresholds. Since both methods have similar mathematical structure, their performance is expected to be comparable. This fact will be verified in Section 6.3.4. However, the method based on target SNR uses less computational resources and is independent of the estimation of the auditory masking thresholds. Hence, this method is preferable for a practical implementation.

To improve speech intelligibility, this chapter introduced a method that uses an ideal binary mask to modify the MWF gains (MWF-IDBM). IDBM is known to improve speech intelligibility but introduce processing artifacts. Since the proposed method is a combination of MWF and IDBM, these processing artifacts are expected to be minimized by the MWF component, while the noise-reduction performance of the MWF component and the speech intelligibility improvement of the IDBM component are expected to be maintained. The latter will be verified in the next chapter (Section 6.3.6). However, when a realistic implementation is considered, i.e., when the mask is estimated from the noisy signal, the ideal speech intelligibility improvement is dramatically reduced.

With respect to computational complexity, Table 2 states that the computational complexity of the FFT-based processing is significantly reduced by the use of an auditory filterbank in the PMWF methods. Figure 17 shows the number of operations required by

Table 2: Comparison between the FFT-based MWF processing and the perceptual MWF processing (PMWF) in its three proposed implementations: IIR filterbank (FB), wavelet packet (WP), and frequency-warped filters (WP).

	FFT-based MWF	**FB-PMWF**	**WP-PMWF**	**FW-PMWF**
Analysis / Synthesis	FFT ($L = 128$) (64 sub-bands)	Auditory filterbank implemented with 4th-order IIR filters (20 sub-bands at 16 kHz, and 24 at 22 kHz)	WP that resembles the auditory filterbank (20 sub-bands at 16 kHz, and 24 at 22 kHz)	frequency-warped filters. Number of sub-bands depends on number of all-pass filters ($K = 16$)
Frequency Resolution	Depends on FFT length. Identical for all T-F bins	Independent on frame length. High at low-frequency and low at high-frequency	Independent on frame length. High at low-frequency and low at high-frequency	Depends on the number of all-pass filters. High at low-frequency and low at high-frequency
Computational Complexity	Depends on FFT length: $L = 128 \rightarrow K = 64$ sub-bands $\rightarrow$ 64 complex linear solvers of size $2M \times 2M$	Fixed. $f_s = 16kHz \rightarrow K = 20$ sub-bands$\rightarrow$20 real linear solvers of size $2M \times 2M$	Depends on WP tree: $f_s = 16kHz \rightarrow K = 20$ sub-bands$\rightarrow$20 real linear solvers of size $2M \times 2M$	Depends on the number of all-pass filters. One real or complex linear solver of size $2MK \times 2MK$
Latency	Fixed. Depends on L. 5.8ms for $L = 128$ and $f_s = 16kHz$	Fixed. One sample if sample-by-sample processing is used for the filterbank	Variable. Depends on the sub-band. 6ms for low-frequency sub-bands and 0.4ms for high-frequency sub-bands	Fixed. One sample if sample-by-sample processing is used for the frequency-warped filters.
Transmission-bandwidth Reduction	Different methods introduced in [13, 94]to reduce the number of channels but still require transmission at full rate for these channels.	Same strategies of FFT-based processing	Multi-rate processing provides bandwidth reduction at different sub-bands. See Section 5.2.2	Not possible

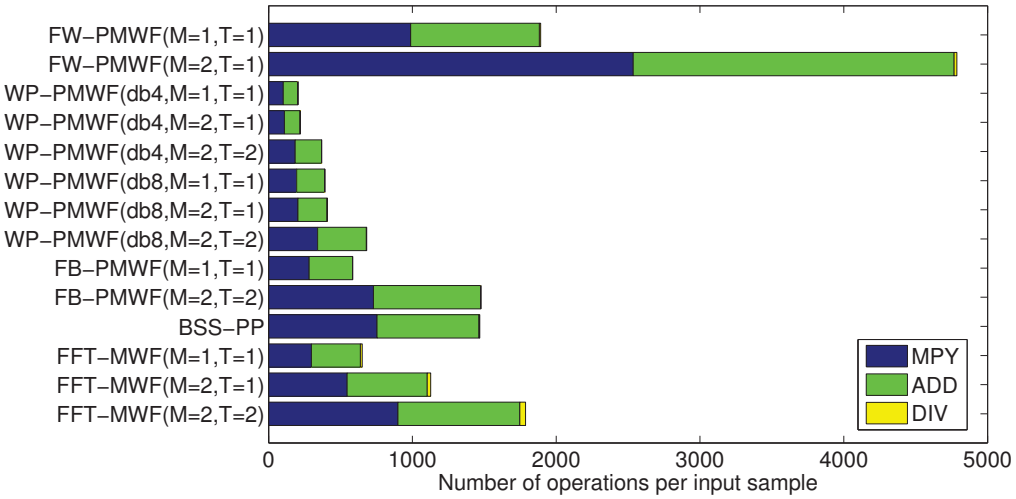

Figure 17: Computational cost of the proposed methods and the FFT-based MWF method for different number of microphones per hearing aid (M) and transmitted channels (T). The cost of FW-PMWF for $M = 2$ and $T = 2$ is not included since it is out of scale.

the proposed and the FFT-based processing to process an input sample at 16 kHz sampling rate, i.e., the total number of operations required to process a frame of length L divided into the frame length. These operations are grouped into multiplications (MPY), additions (ADD), and divisions (DIV). For comparison purposes, the number of operations in the BSS-PP method are also included in this plot. This figure shows how the number of operations in an FFT-based processing is reduced dramatically (around 70%) by using the proposed WP-PMWF method with mother wavelet db4. On the contrary, the number of operations in the FB and FW implementations is not reduced with respect to the FFT-based implementation. This fact supports the claim that the WP-PMWF method provides significant advantages over the other PMWF and the FFT-based MWF methods.

It is interesting to show that in the FFT-based processing, the bottleneck corresponds to the estimation of the second-order statistics and the weight computation (Fig. 18). However, in WP-PMWF the number of operations required for these stages is insignificant compared to the number of operations required for the computation of the auditory representation. This results suggests that the WP-PMWF method can be accelerated by hardware components dedicated to the computation of the WP tree. The Chapter 7 explores other alternatives to reduce the bottleneck in the statistics update and weight computation

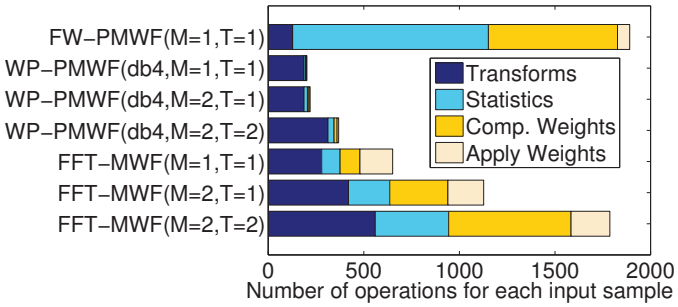

Figure 18: Computational cost of PMWF and FFT-based MWF reported for each functional group and different number of microphones per hearing aid (M) and transmitted channels (T).

on the FFT-based implementation and the FW-PMWF method, and the reduction of the bottleneck in the WP-PMWF processing due to the WP computation.

Chapter VI

PERFORMANCE EVALUATION OF THE PROPOSED METHODS

To reduce computational cost and latency, this research proposes two perceptually-inspired methods. The first method is a BSS-based binaural noise-reduction method that uses a BSS algorithm to get estimates for the speech and noise signals, and these estimates are used in a perceptually-inspired post processing to compute the time-domain gains that cancel out the background noise. This method, called blind source separation with perceptual post processing (BSS-PP), was introduced in the Chapter 4. The second method is based on MWF. In this case, the FFT has been replaced by an auditory filterbank. The proposed MWF method, called perceptual MWF (PMWF), was introduced in the Chapter 5. Three implementations were introduced for auditory filterbank: IIR filterbank (FB), wavelet packet (WP), and frequency-warped filters (FW). This auditory processing reduces the number of sub-bands and increases the frequency resolution for the low-frequency sub-bands compared to an FFT-based MWF. In addition, the processing involves exclusively real operations instead of complex operations as in the case of the FFT-based MWF, which reduces the computational cost significantly as shown in Section 5. Chapter 5 also introduced different implementation strategies to estimate the second-order statistics (MWF-CPSDμ_{SNR}) and to improve the speech intelligibility (MWF-IDBM).

This chapter discusses the performance of all above methods: Section 6.2 the performance of the BSS-PP method, Sections 6.3 and 6.4, the performance of the different variants of the PMWF method. A final comparison of BSS-PP and PMWF is presented in Section 6.5.

6.1 Experiment

The methods proposed in this research, BSS-PP (Section 4.2), PMWF (Section 5.2), MWF-CPSDμ_{SNR} (Section 5.3), MWF-μ_{ATH} (Section 5.4), and MWF-IDBM (Section 5.5) are evaluated using the database of the comparative study in Chapter 3. Since this database

is for non-reverberant speech, a secondary database using reverberant conditions is created using the HRTF recordings described in [29, 84]. This database is included since it is widely known that the performance of the majority of the noise-reduction algorithms is degraded significantly when reverberation is present. This database assumes a babble noise scenario and the following rooms: studio ($RT_{60} = 0.12$s), meeting room ($RT_{60} = 0.23$s), office ($RT_{60} = 0.43$s), and lecture room ($RT_{60} = 0.78$s).

The performance evaluation is performed by objective metrics and subjective tests. The objective metrics used in this study are: the broadband intelligibility weighted SNR improvement (ΔSNR-SII) [21], the noise power level reduction (NPLR) [18], perceptual evaluation of speech quality (PESQ) [27], and the coherence-based intelligibility index (I3) [34]. The values for ΔSNR-SII, NPLR, PESQ, and I3 reported in this chapter correspond to the average over 10 different speech utterances.

For the subjective test, MOS (mean opinion score) and MUSHRA (multiple stimulus test with hidden reference and anchor) are used to assess the overall sound quality. For MOS tests, significant differences are assessed by analysis of variance (ANOVA). The protocols in [28, 26] are used for these subjective tests. All subjective tests conducted in this document are performed with normal-hearing listeners. Since hearing impairments vary widely, performing subjective tests for all kinds of hearing impairments is infeasible in this study. Hearing-impaired listeners are expected to perceive less amount of processing artifacts than normal listeners but each hearing-impaired listener may perceive different artifacts to different degrees. Therefore, by using normal listeners in the subjective test, we are able to evaluate the full range of artifacts that may be perceived by listeners having various impairments.

6.2 Performance of Perceptually Inspired BSS-Based Method

For comparison purposes, the performance of the proposed method, BSS-PP, is compared to three existing methods: BSS-Aichner-07, BSS-Reindl-10, and MWF-N. SNR improvement for diffusive, babble, and multi-talker scenarios is plotted in Figures 19-21. In general, the proposed method (BSS-PP) outperforms the existing methods in most scenarios.

The poor performance of the proposed method in the multi-talker scenario at low input SNR is explained by the errors introduced by a wrong selection of the primary output. When an ideal output selection algorithm is used (dashed line in Figure 21), the performance of BSS-PP is similar or better than that of the existing methods. The output selection algorithm can be made more robust by using a direction-of-arrival-estimation algorithm or a permutation algorithm at expenses of increasing the computational complexity. However, scenarios with very few interfering signals at input SNR < 0 dB such as the multi-talker scenario of Fig. 21 are very uncommon, and they are not challenging for the auditory system without any hearing aid. Likewise, binaural noise-reduction methods are useful for challenging scenarios such as babble noise at low input SNR. Since the proposed method provides an excellent performance under these scenarios (Fig. 20), the output-selection algorithm used by our method is enough for a large set of practical applications.

Up to this point the performance of BSS-PP has been verified under non-reverberant scenarios. For reverberant scenarios, Figures 22-24 show that the proposed method provides an acceptable SNR improvement (>3 dB) for rooms with low and large reverberation. In addition, the proposed method outperforms the existing methods for an input SNR≥ 0 dB. The poor performance of the proposed method for input SNR <0 dB is explained by errors in the source separation performed by the BSS algorithm.

The results for the MOS subjective test are summarized in the Table 3. Quality and noise reduction are graded in the scale $[0, 5]$, with 5 the highest value. Percentages in the preservation of binaural cues correspond to the number of subjects who identified correctly the direction of arrival. Subjects showed the preference for the proposed method regarding the noise reduction efficiency. However, the quality of the proposed method is lower than MWF-N but higher than BSS-Aichner-07. Moreover, the proposed method preserves the localization cues for both target and interfering signals. This feature is an advantage over the BSS-Aichner-07 method, in which noise localization cues are not preserved. We derived mathematical proofs about the preservation of localization cues for BSS-PP and BSS-Aichner-07 methods in [63], concluding that BSS-PP can preserve localization cues for both target and interfering signals simultaneously. However, BSS-Aichner-07 can preserve

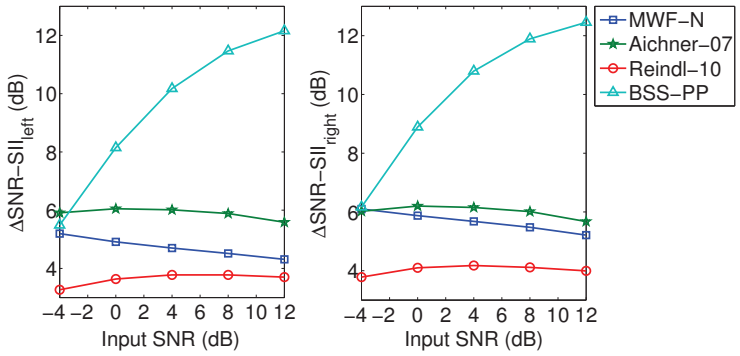

Figure 19: SNR improvement for BSS-PP under diffusive noise scenario.

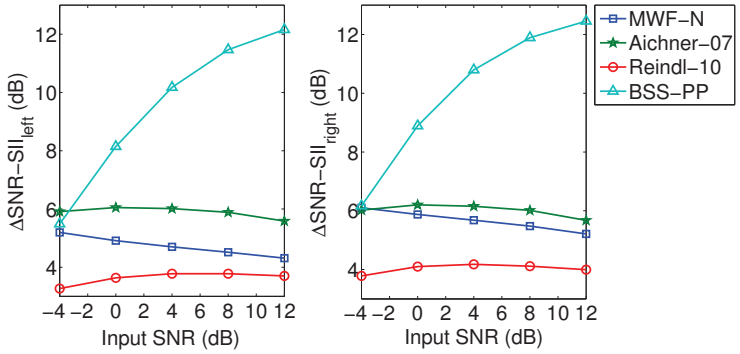

Figure 20: SNR improvement for BSS-PP under babble noise scenario.

both localization cues only in the determined case, i.e., when the number of interfering signals is less than the number of microphones.

6.3 Performance of Perceptually Inspired MWF-Based Method

This section discusses the performance evaluation of the PMWF method using different objective metrics. For all experiments in this section, the MWF weights are computed using the SDW-MWF framework, i.e., through (18) and (19).

Table 3: Subjective test results for the proposed method (BSS-PP), the BSS method proposed by Aichner, and MWF-N.

	Clean Signal	Original Signal	BSS Aichner-07	BSS-PP	MWF-N
Speech Quality	4.7	4.0	2.4	3.0	3.8
Noise Reduction	4.8	1.9	2.1	2.6	2.3
Target Cues	97%	90%	83%	83%	82%
Noise Cues	96%	97%	58%	85%	79%

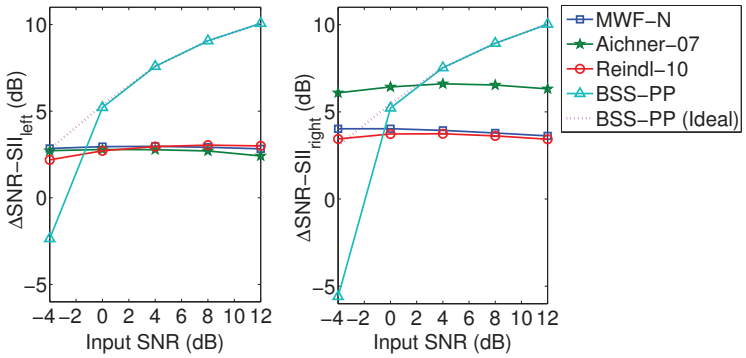

Figure 21: SNR improvement for BSS-PP under multi-talker scenario. The dashed line is the performance for an ideal output-selection algorithm.

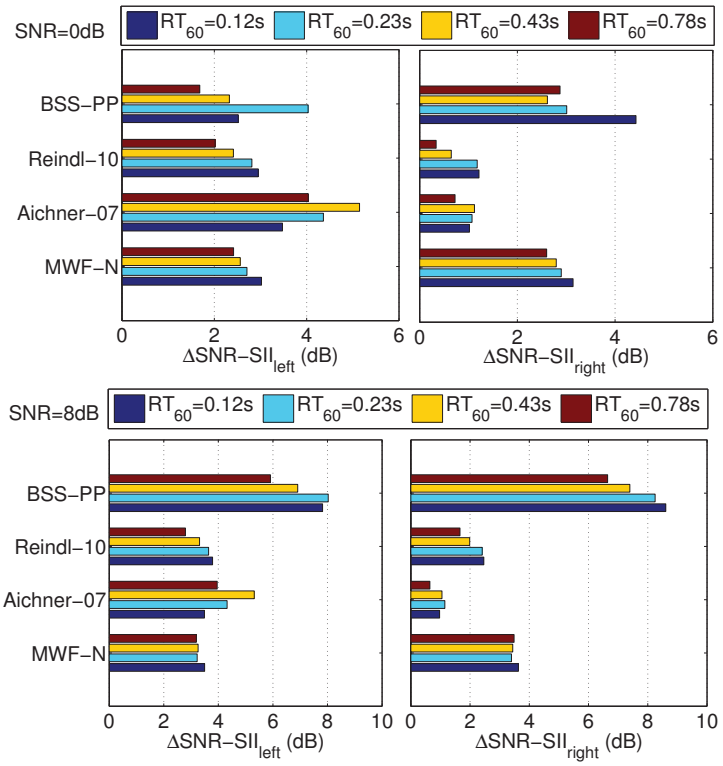

Figure 22: SNR improvement for BSS-PP under babble noise scenario in different reverberant rooms.

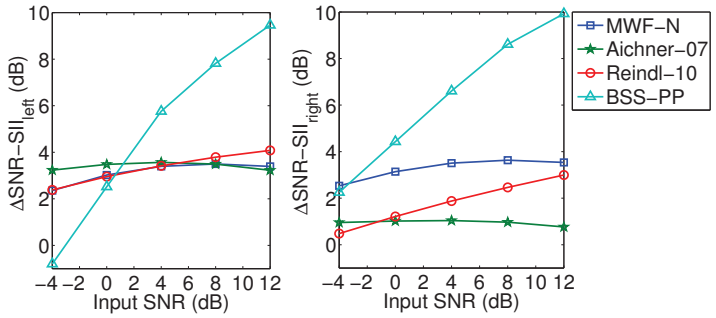

Figure 23: SNR improvement for BSS-PP under babble noise scenario in a studio room (reverberant condition $RT_{60} = 0.12$s).

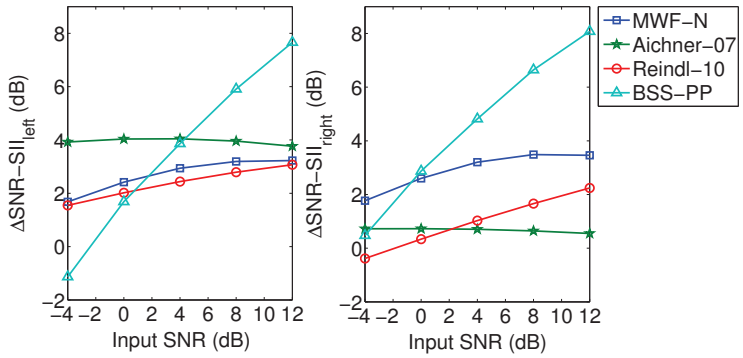

Figure 24: SNR improvement for BSS-PP under babble noise scenario in a lecture room (reverberant condition $RT_{60} = 0.78$s).

6.3.1 MWF Using Auditory Filterbank Based on Wavelet Packet (WP-PMWF)

In Section 5.2, three methods to implement the auditory transformation were mentioned, an IIR filterbank (FB), wavelet packet (WP) filterbank, and frequency-warped filters (FW). In [55], we compared the FB and WP approaches, concluding that both implementations provide similar performance in terms of SNR improvement and sound quality, but the WP-based approach employs less computational resources. The following discussions are focused exclusively on WP-based PMWF (WP-PMWF). The results for the FW-based approach are discussed in the next section.

The performance of the proposed method depends on two parameters, the trade-off parameter μ and the frame length L. In [54], we investigated the effect of these parameters under diffusive and babble noise scenarios. The following conclusions are derived in [54]:

- **Effect of the trade-off parameter** μ. The SNR improvement is increased with the use of a large μ, as expected, and a large μ also provides better sound quality. This improvement on the sound quality is a result of the higher noise reduction.

- **Effect of the frame length** L. The proposed processing requires small frame length to achieve higher SNR improvement and sound quality. This performance can be explained by the independence of the number of sub-bands on the frame length L, and the way how the statistics are updated. In the FFT-based processing, the low-frequency resolution depends on the frame length L, but in the proposed approach the number of sub-bands is fixed and independent on L, so that the frequency resolution is expected to be high at low frequencies regardless the value of L. In addition, when L is small, the second-order statistics estimators can track rapid changes in speech and noise statistics, providing better estimation, and so better noise reduction. The fact that PMWF requires small frame lengths is an interesting result for a real-time implementation since using short L involves reduction in the latency.

To compare the performance of WP-PMWF and FFT-based MWF, both methods are tested under babble, small car, and street noise scenarios. The second-order statistics are estimated

assuming a perfect VAD to obtain the upper-bound performance of both algorithms. The performance of WP-PMWF using the second-order statistics estimation based on CPSD (MWF-CPSDμ_{SNR}) is presented in the Section 6.3.3. The FFT-based implementation is tested for $\mu = 5$, which is a value reported to provide good SNR improvement and low speech distortion [13]. In addition, two FFT lengths are tested, $L = 128$ and $L = 32$. $L = 128$ is reported to provide good sound quality and SNR improvement [13]. However, the usage of $L = 32$ is known to introduce large speech distortion. The implementation for $L = 32$ is considered since this is a common FFT length in some commercial digital hearing aid devices, and the number of sub-bands (16) is comparable to the number of sub-bands in PMWF (20 at $f_s = 16kHz$). The WP-based implementation uses mother wavelet Daubechies 8 (db8). The effect of different wavelet families is shown later. Two objective metrics are employed for this analysis, the weighted SNR improvement (ΔSNR-SII) and the noise power level reduction (NPLR).

The SNR improvement and the noise reduction in the proposed method are improved by increasing μ as expected (Figures 25, 27, and 29). In average, the proposed method outperforms the FFT-based processing (FFT-MWF) for all scenarios analyzed. The SNR improvement of the proposed method for $\mu \geq 10$ is comparable to the SNR improvement of the FFT-based processing (Figures 25, 27, and 29). Moreover, the noise reduction is more significant in the proposed method at low-input SNR (Figures 26, 28, and 30). This result is explained by the higher low-frequency resolution provided by the PMWF method compared to the FFT-based implementation. Whereas the FFT-based implementation using small FFT length ($L = 32$) provides the poorest SNR improvement and noise reduction, the proposed method using similar number of sub-bands provides the best SNR improvement and noise reduction. This result shows the significant advantages of WP-PMWF over the FFT-based method.

The proposed implementation allows the usage of different mother wavelets. The computational complexity of the proposed method depends strongly on the WP tree and the order of the mother wavelet. For example, a WP implemented with Daubechies 4 involves one half the number of operations of a Daubechies 8. To identify the effect of the mother

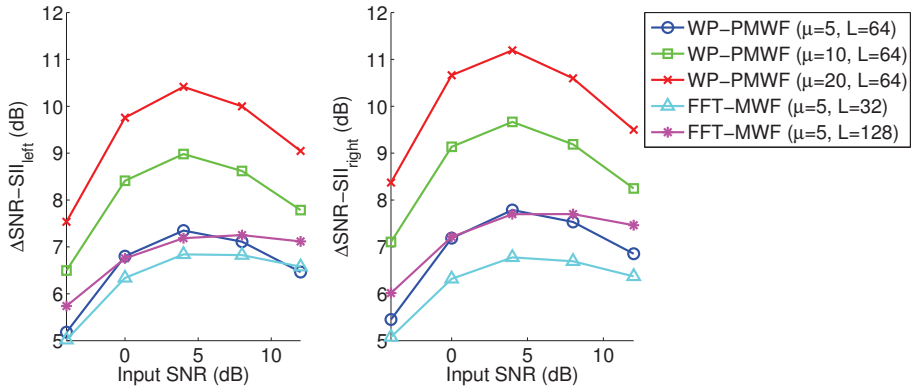

Figure 25: SNR improvement for the perceptually-based processing (WP-PMWF) and the FFT-based processing (FFT-MWF) under babble noise scenario.

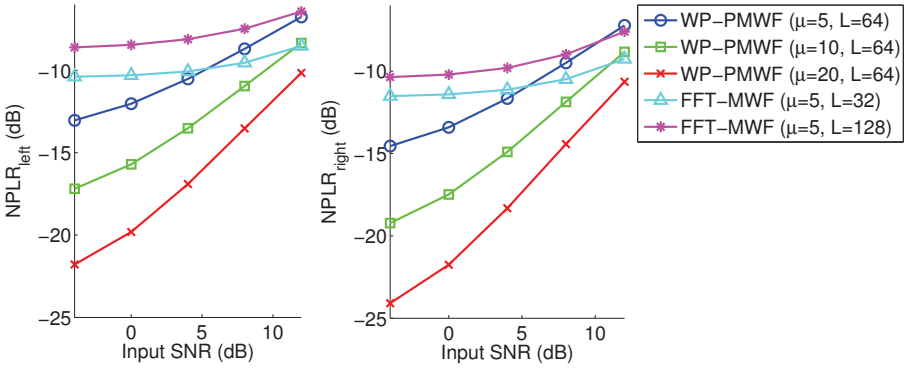

Figure 26: Noise reduction for the perceptually-based processing (WP-PMWF) and the FFT-based processing (FFT-MWF) under babble noise scenario.

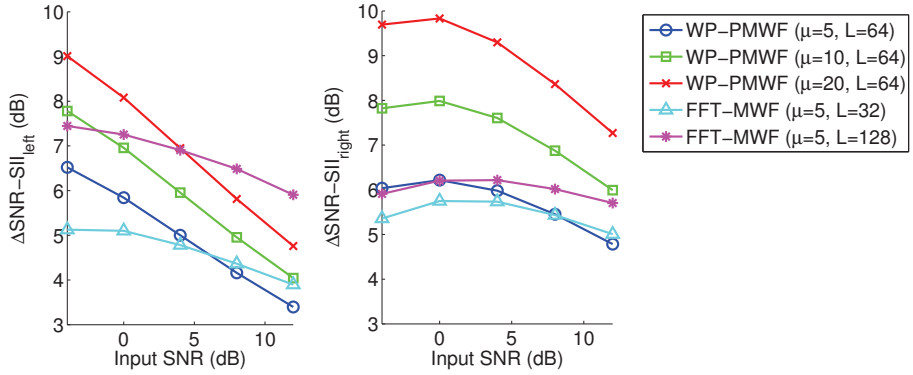

Figure 27: SNR improvement for the perceptually-based processing (WP-PMWF) and the FFT-based processing (FFT-MWF) under small car noise scenario.

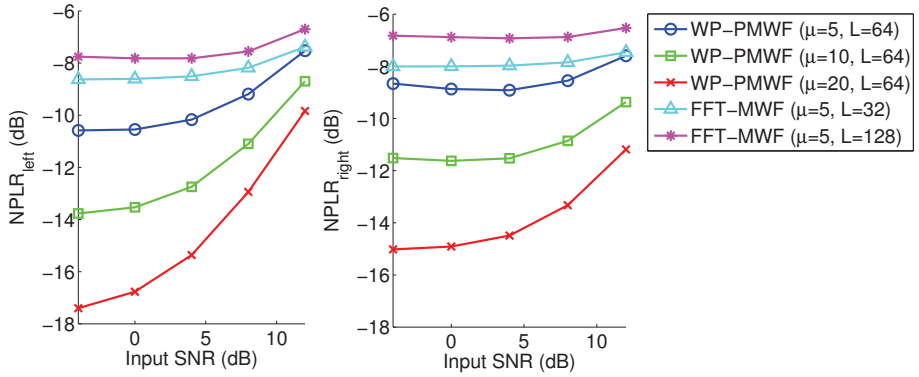

Figure 28: Noise reduction for the perceptually-based processing (WP-PMWF) and the FFT-based processing (FFT-MWF) under small car noise scenario.

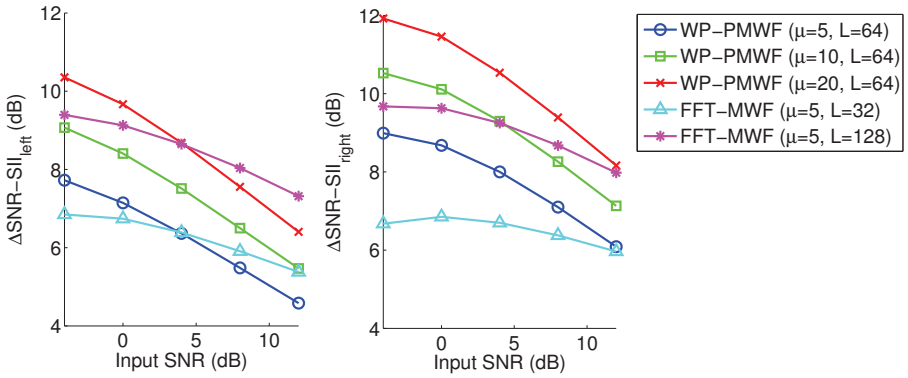

Figure 29: SNR improvement for the perceptually-based processing (WP-PMWF) and the FFT-based processing (FFT-MWF) under street noise scenario.

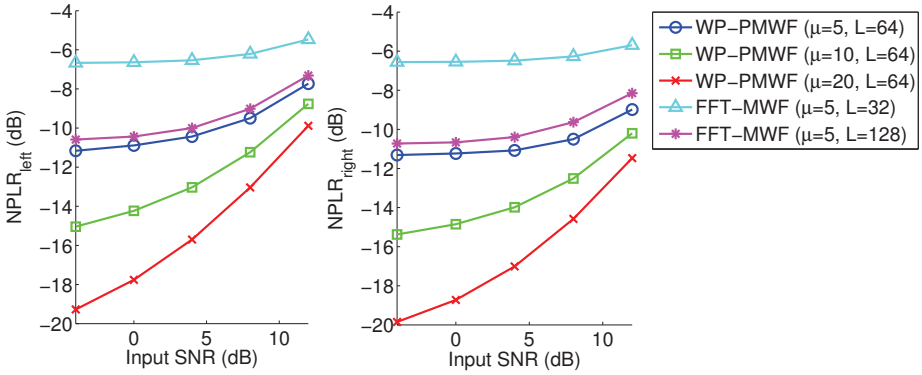

Figure 30: Noise reduction for the perceptually-based processing (WP-PMWF) and the FFT-based processing (FFT-MWF) under street noise scenario.

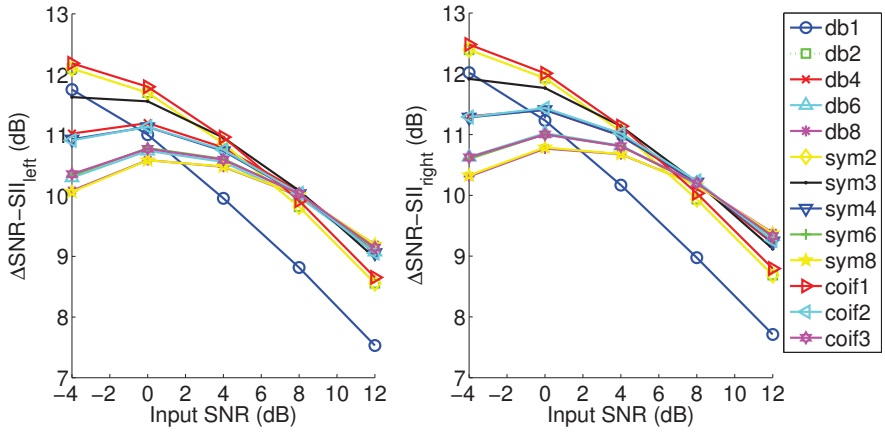

Figure 31: SNR improvement for WP-PMWF implemented with different mother wavelet (Daubechies–db, Symlets–sym, and Coiflets–coif) under diffusive noise scenario.

wavelet on the performance of the proposed method, simulations for diffusive and babble noise scenarios are carried out for the following wavelet families: Daubechies (orders 1, 2, 4, 6, and 8), Symlets (orders 2, 3, 4, 6, and 8), and Coiflets (orders 1, 2, and 3). Figures 31-36 show the SNR improvement, noise reduction, and objective quality (PESQ) for different mother wavelets. For a given wavelet family, increasing the order reduces the SNR improvement and the noise reduction but improves speech quality. The exception is the diffusive noise scenario in which the noise reduction is independent on the mother wavelet.

High-order mother wavelets are preferable to meet speech quality requirements. However, high-order mother wavelets degrade the SNR improvement and noise reduction, and increase the computational cost. Hence, it is preferable to keep the wavelet order small enough to achieve good noise reduction and computational cost. Particularly, the Daubechies wavelets for orders $n \geq 4$, Symlet wavelets for orders $n \geq 4$, and Coiflet wavelets for orders $n \geq 2$ provide acceptable sound quality and good noise reduction. Informal listening tests show no sound quality differences between db4 and db8, or db8 and sym4. For this reason, db4 is selected as mother wavelet for all further experiments.

6.3.2 MWF Using Frequency-Warped Filters (FW-PMWF)

In the Section 5.2.3, two approaches based on frequency-warped filters (FW) were introduced as an alternative to implement PMWF. Both approaches differ on the information

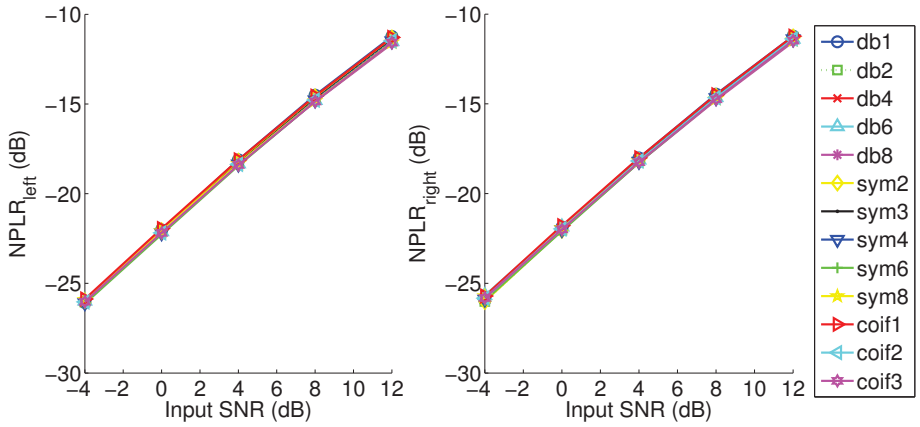

Figure 32: Noise reduction (NPLR) for WP-PMWF implemented with different mother wavelet (Daubechies–db, Symlets–sym, and Coiflets–coif) under diffusive noise scenario.

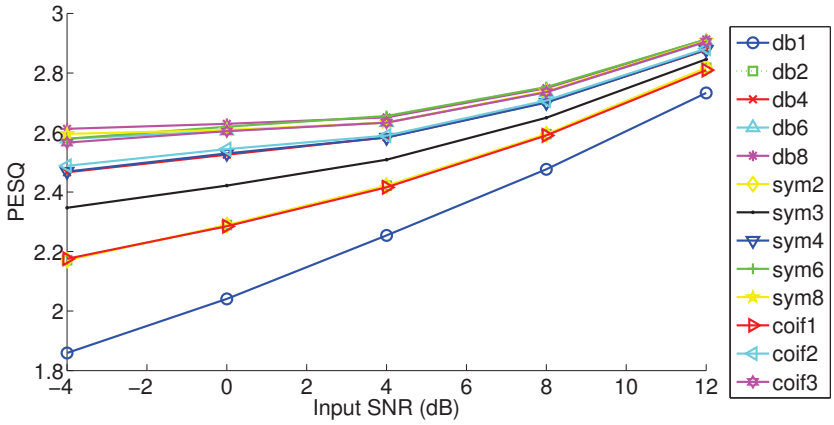

Figure 33: Objective quality (PESQ) for WP-PMWF implemented with different mother wavelet (Daubechies–db, Symlets–sym, and Coiflets–coif) under diffusive noise scenario.

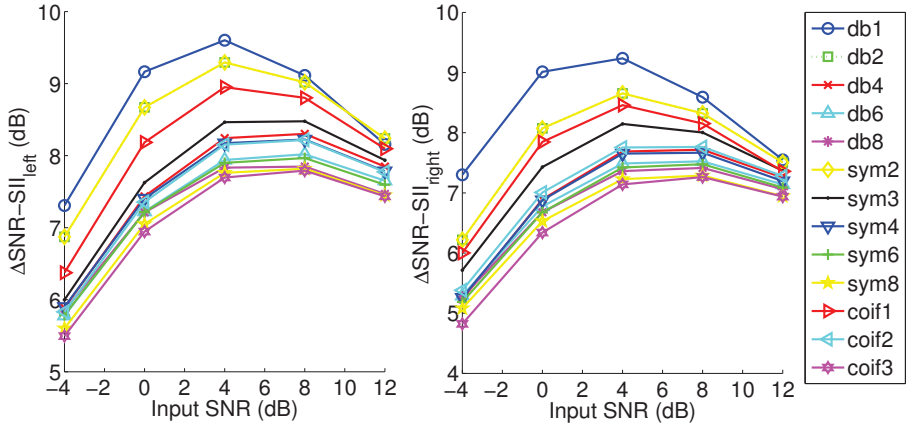

Figure 34: SNR improvement for WP-PMWF implemented with different mother wavelet (Daubechies–db, Symlets–sym, and Coiflets–coif) under babble noise scenario.

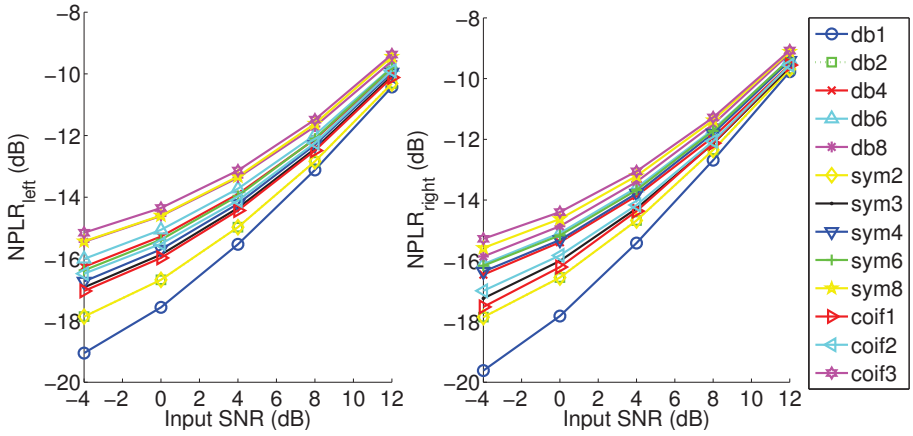

Figure 35: Noise reduction (NPLR) for WP-PMWF implemented with different mother wavelet (Daubechies–db, Symlets–sym, and Coiflets–coif) under babble noise scenario.

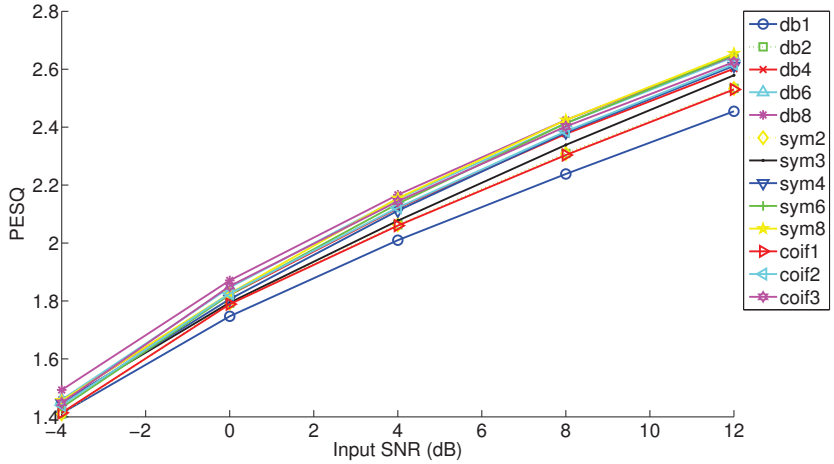

Figure 36: Objective quality (PESQ) for WP-PMWF implemented with different mother wavelet (Daubechies–db, Symlets–sym, and Coiflets–coif) under babble noise scenario.

used to estimate the weights. If this information is the output of the all-pass filters, the method is referred as the time-warped FW-PMWF. On the other hand, if this information is the output of the FFT taken at the all-pass filters, the method is called frequency-warped FW-PMWF. The performance of these FW-PMWF methods is presented in the Figures 37-39. The performance of WP-PMWF is also included in these plots for comparison purposes. All plots are generated for a trade-off parameter $\mu = 10$, and the number of taps in the all-pass filter chain is 16. For all metrics analyzed, SNR improvement, noise reduction, and objective quality, the time-warped FW-PMWF method outperforms the frequency-warped FW-PMWF method, and the performance of the time-warped FW-PMWF method is nearly to that of the WP-PMWF method. In terms of computational cost, time-warped FW-PMWF offers less computational cost than frequency-warped FW-PMWF. However, the computational cost of FW-PMWF is higher than the WP-PMWF method (Section 5.6), which supports the claim about the significant advantages of the WP-based implementation. For this reason, the WP-PMWF is identified as more promising for a real-time binaural hearing aid.

6.3.3 Effect of Non-VAD Second-Order Statistics Estimation

To verify the performance of the second-order statistics estimation method based on CPSD and adaptive μ (MWF-CPSDμ_{SNR}, Section 5.3), the SDW-MWF method is implemented initially using FFT processing (FFT length $L = 128$) and the following approaches [56]:

1. Update of the statistics using perfect VAD, first-order estimator (20), and fixed $\mu = 5$ (PVADμ_{fix}). A perfect VAD is used to estimate the upper-bound performance of the VAD-based methods.

2. Framework described by Ngo *et al.* [67]. (RVADμ_{prob}). This framework updates the statistics using a soft-VAD and adaptive μ.

3. Update of the statistics using the proposed CPSD estimator and an adaptive μ (38). (CPSDμ_{SNR}). The following parameters are used for all experiments: $\alpha_v = 0.9$, $\alpha_x = 0.97$, $\delta = 0.9$, and $\epsilon = 3.0$.

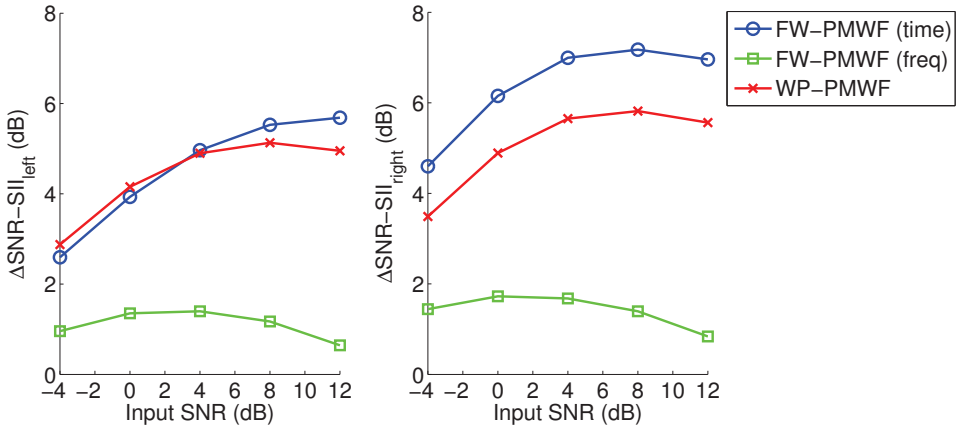

Figure 37: SNR improvement for PMWF implemented with frequency-warped filters under babble noise scenario.

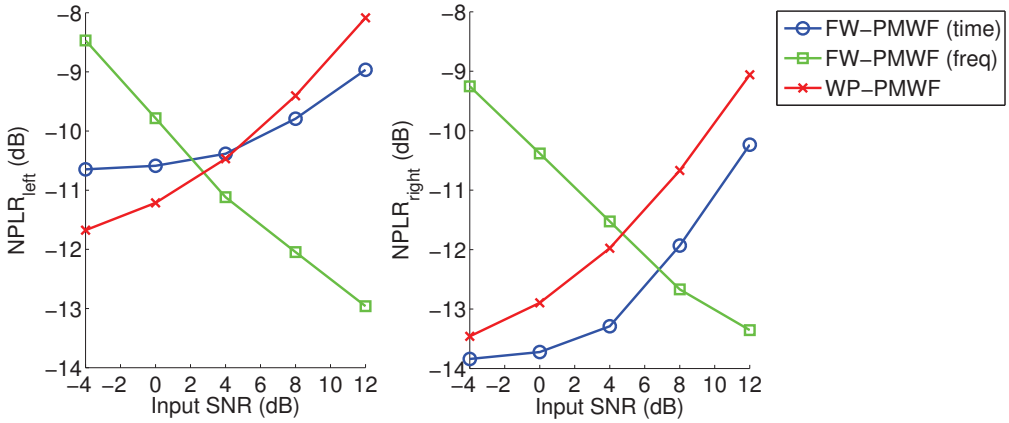

Figure 38: Noise reduction (NPLR) for PMWF implemented with frequency-warped filters under babble noise scenario.

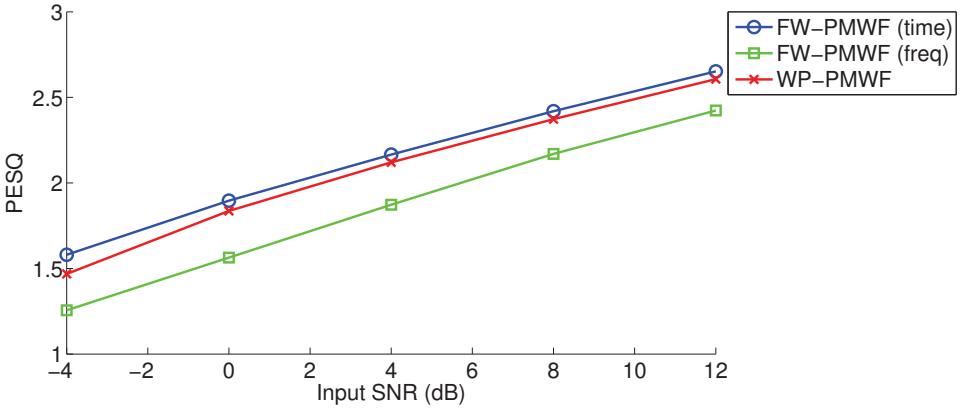

Figure 39: Objective quality (PESQ) for PMWF implemented with frequency-warped filters under babble noise scenario.

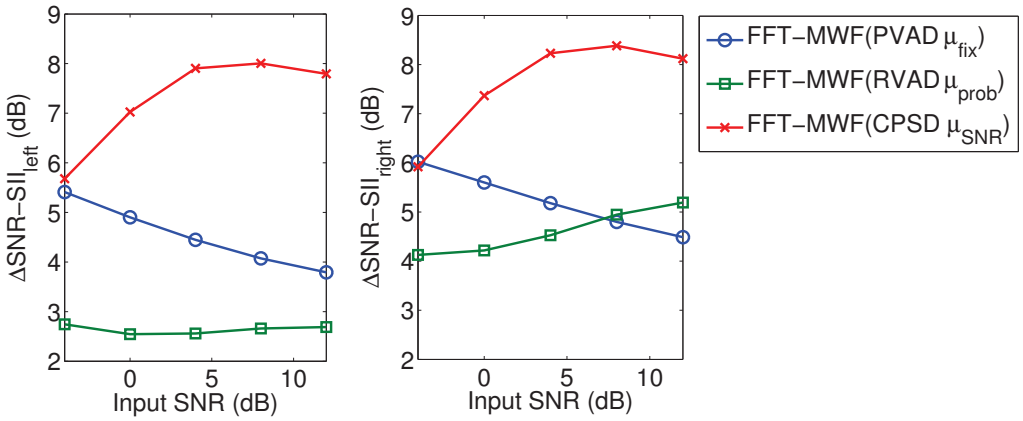

Figure 40: SNR improvement of MWF-CPSDμ_{SNR} under constant-SNR babble noise.

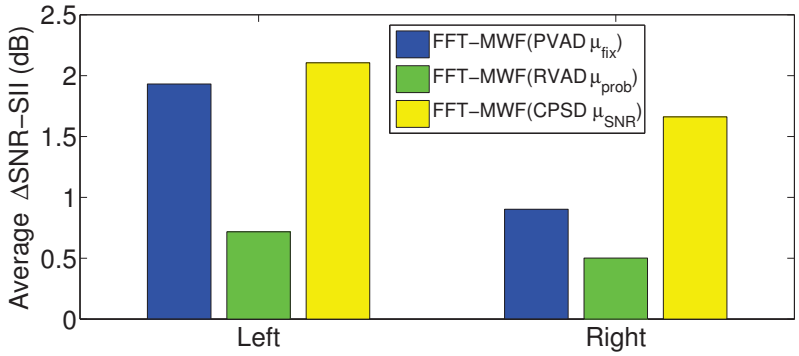

Figure 41: SNR improvement of MWF-CPSDμ_{SNR} under variant-SNR babble noise.

For all cases, the statistics are updated in the frame-by-frame basis. We analyzed two scenarios in [56], constant-SNR babble noise (inside a cafeteria) and variant-SNR babble noise (getting in a cafeteria).

In overall, the proposed implementation (CPSDμ_{SNR}) provides better SNR improvement than the VAD-based approaches (Fig. 40-41). As we expected, a VAD-based estimation, even with perfect VAD, is unable to provide significant SNR improvement for highly non-stationary environments such as the variant-SNR babble noise scenario (Fig. 41). This limitation is overcome with the proposed implementation based on CPSD.

The sound quality of the CPSDμ_{SNR} implementation is slightly above the perfect-VAD and fixed-μ implementation (PVADμ_{fix}) (Fig. 42). The latter suggests that the proposed method improves the SNR and preserves the sound quality simultaneously. This property is explained by the adaptive μ. In the proposed method, μ is adapted to meet a desired SNR

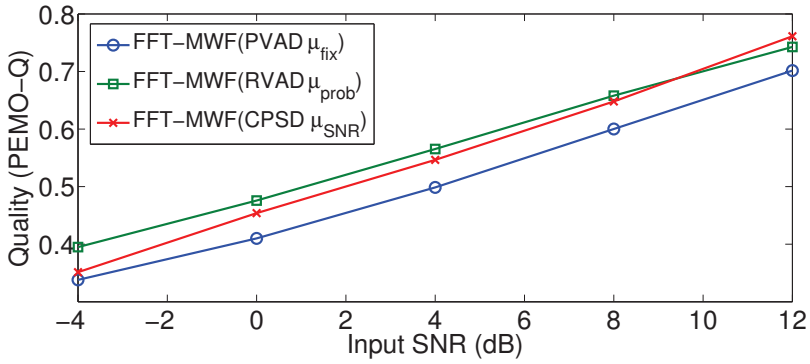

Figure 42: Objective quality of MWF-CPSDμ_{SNR} under constant-SNR babble noise.

for each frequency bin, and the values of μ provided by (38) lead to small speech distortion. A different situation is present in the method by Ngo *et al.* (RVADμ_{prob}). In the Ngo's method, μ is adapted according to the probability of being a voiced segment. Thus, small μ is used to process speech segments, and large μ for noise-only segments. But this adaptive μ is not constrained to the noise level as in the proposed method. For this reason, RVADμ_{prob} provides high sound quality even though the SNR improvement is not as significant as for PVADμ_{fix} or CPSDμ_{SNR}.

Up to this point, the performance of the CPSD-based estimation has been discussed for the FFT-based implementation of SDW-MWF showing significant benefits under highly non-stationary environments. In addition, WP-PMWF showed significant advantages over the FFT-based implementation. But the performance of WP-PMWF presented in the Section 6.3.1 assumed perfect VAD to estimate the second-order statistics. When the CPSD-based estimation is used for WP-PMWF, this estimation strategy also provides significant benefits[1]. In particular, the SNR improvement of the CPSD-based estimation is comparable to VAD-based estimation using perfect VAD and high μ ($\mu = 20$) (Figure 43); the noise reduction is more aggressive with a CPSD-based estimation (Figure 44); and the sound quality is not degraded significantly (Figure 45).

[1]The following parameters are selected for the CPSD-based estimation algorithm: $\alpha_v = 0.97$, $\alpha_x = 0.99$, $\delta = 0.9$, and $\epsilon = 3.0$

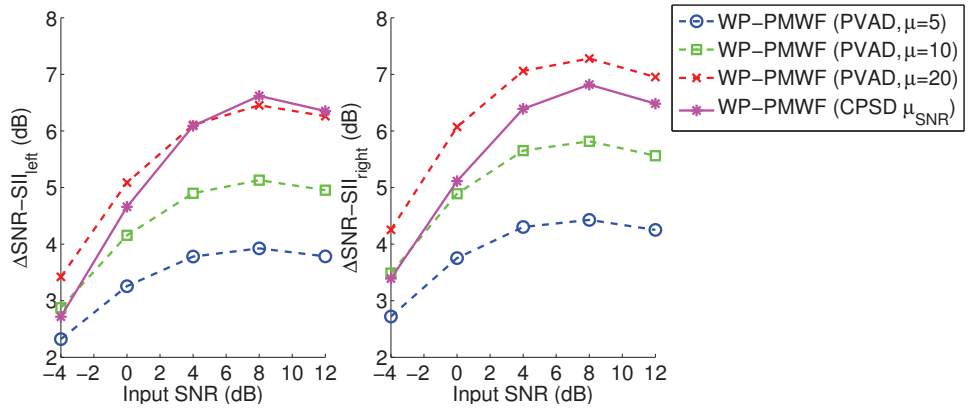

Figure 43: SNR improvement of MWF-CPSDμ_{SNR} under babble noise scenario. All plots are for WP-PMWF implemented with two strategies to update the statistics: perfect VAD and CPSD.

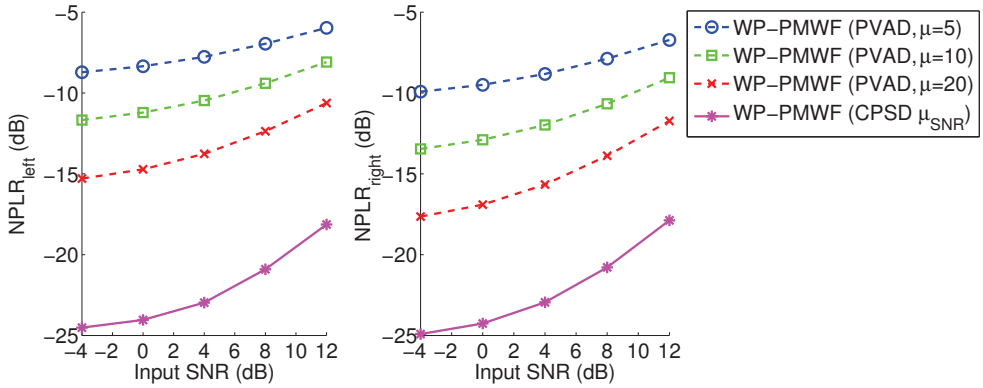

Figure 44: Noise reduction of MWF-CPSDμ_{SNR} under babble noise scenario. All plots are for PMWF implemented with two strategies to update the statistics: perfect VAD and CPSD.

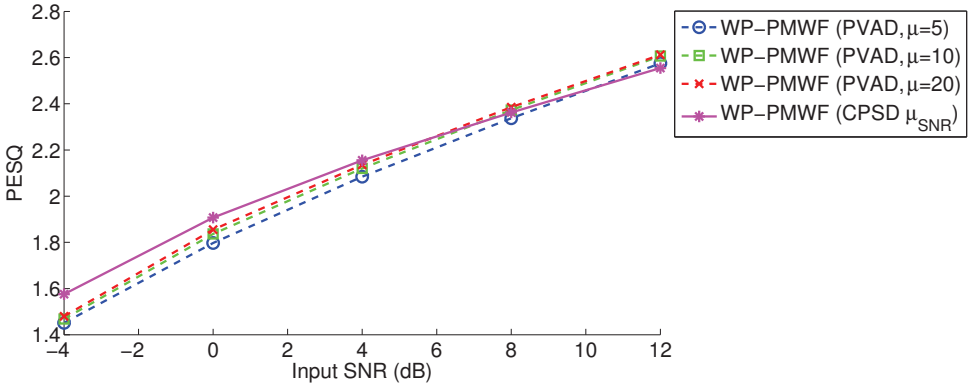

Figure 45: Objective quality of MWF-CPSDμ_{SNR} under babble noise scenario. All plots are for PMWF implemented with two strategies to update the statistics: perfect VAD and CPSD.

6.3.4 MWF Framework Based on Auditory Masking Threshold

Results showed in the previous sections correspond to the SDW-MWF framework implemented with two strategies: replacement of the FFTs by an auditory filterbank using wavelet packet (WP-PMWF); and second-order statistics estimated by a CPSD-based estimator and adaptive trade-off parameter μ based on target SNR (MWF-CPSDμ_{SNR}). In Section 5.4, a MWF framework based on auditory masking thresholds (MWF-μ_{ATH}) was derived. For this framework, the equations to compute the weights are the same as for SDW-MWF except that the trade-off parameter μ is adapted according to the auditory masking thresholds.

Initial simulations of MWF-μ_{ATH} are conducted for an FFT-based implementation. These simulations considered an FFT length of $L = 128$, and are conducted for three scenarios: one constant-SNR babble noise scenario and two variant-SNR babble noise scenarios (getting in and getting out a cafeteria). Auditory masking thresholds are estimated using the algorithm proposed in [30]. The computation of the auditory masking thresholds requires the estimation of speech and noise power levels. These power levels are extracted from the second-order statistics $\boldsymbol{R}_x$ and $\boldsymbol{R}_v$, respectively. Therefore, the performance of MWF-μ_{ATH} is expected to depend strongly on the estimation of the second-order statistics. To identify the robustness of MWF-μ_{ATH} against estimation errors, two versions of the proposed method are tested: using perfect VAD (PVADμ_{ATH}) and using a real VAD (RVADμ_{ATH}). The performance obtained by PVADμ_{ATH} corresponds to the upper-bound performance for a VAD-based implementation using auditory masking thresholds. The real VAD used in the experiments is taken from the Voicebox's toolbox [9]. The performance of MWF-μ_{ATH} is compared to a SDW-MWF method implemented with two strategies: VAD-based statistics estimation using perfect VAD and fixed $\mu = 5$ (PVADμ_{fix}), and CPSD-based statistics estimation and adaptive μ based on target SNR (CPSDμ_{SNR}).

In average, MWF-μ_{ATH} is a promising method since the performance under all scenarios using real VAD is similar or better to a SDW-MWF method using fixed μ and an unrealistic "perfect VAD" (Figures 46-47). Moreover, there is a strong influence of the estimation errors on the performance of MWF-μ_{ATH} because the performance using perfect VAD (PVADμ_{ATH}) differs in more than 2 dB the performance using real VAD (RVADμ_{ATH}). It

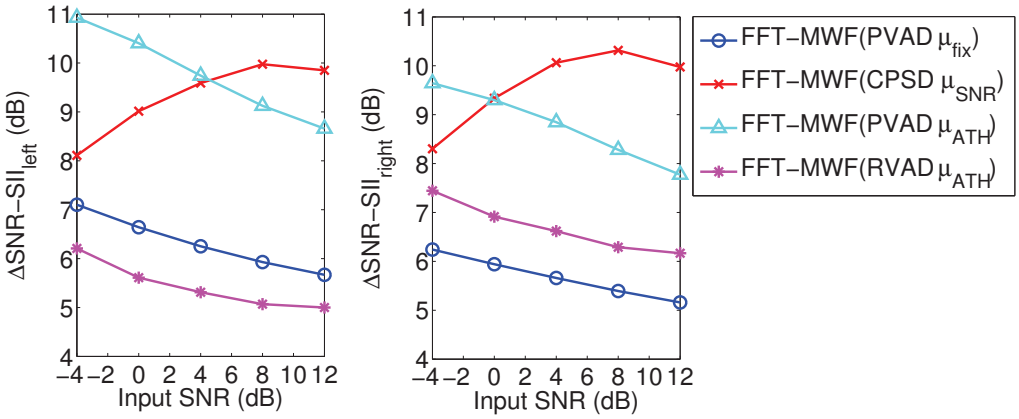

Figure 46: SNR improvement of FFT-MWF implemented with different strategies to estimate the statistics: VAD-based estimation and fixed μ (PVADμ_{fix}), CPSD-based estimation and adaptive μ (MWF-CPSDμ_{SNR}), and MWF framework based on auditory masking thresholds (MWF-μ_{ATH}). All implementations use an FFT length $L = 128$ and the SDW-MWF framework to compute the weights. Scenario: Babble noise.

is important to remark that MWF-CPSDμ_{SNR} provides a performance similar to the upper-bound performance of MWF-μ_{ATH}, i.e., the performance of PVADμ_{ATH}. The similarity in the upper-bound performance of MWF-μ_{SNR} and MWF-CPSDμ_{SNR} is not surprising since both methods were shown to have a similar structure to compute μ. Whereas MWF-μ_{ATH} requires estimates of the speech power and auditory masking threshold to compute μ, MWF-CPSDμ_{SNR} replaces these quantities by a fixed target SNR, which reduces the computational complexity involved in the estimation of the auditory masking threshold. Hence, for practical implementations, MWF-CPSDμ_{SNR} is an ideal substitute to MWF-μ_{ATH}, obtaining similar performance but using fewer operations.

Up to this point the performance of MWF-μ_{ATH} has been tested for an FFT-based processing. For PMWF, simulations of MWF-μ_{ATH} are conducted only for babble noise scenario. The PMWF method is implemented using 4 strategies: a VAD-based estimation using perfect VAD and $\mu = 10$; MWF-CPSDμ_{SNR}; and MWF-μ_{ATH} using perfect VAD and real VAD. The results are similar to those obtained for the FFT-based processing. In terms of SNR improvement, the performance of MWF-μ_{ATH} using real estimates is similar to the performance of an idealistic PMWF implemented with perfect VAD. Besides, the performance of MWF-μ_{ATH} is degraded when real VAD is used. Also, MWF-μ_{ATH}

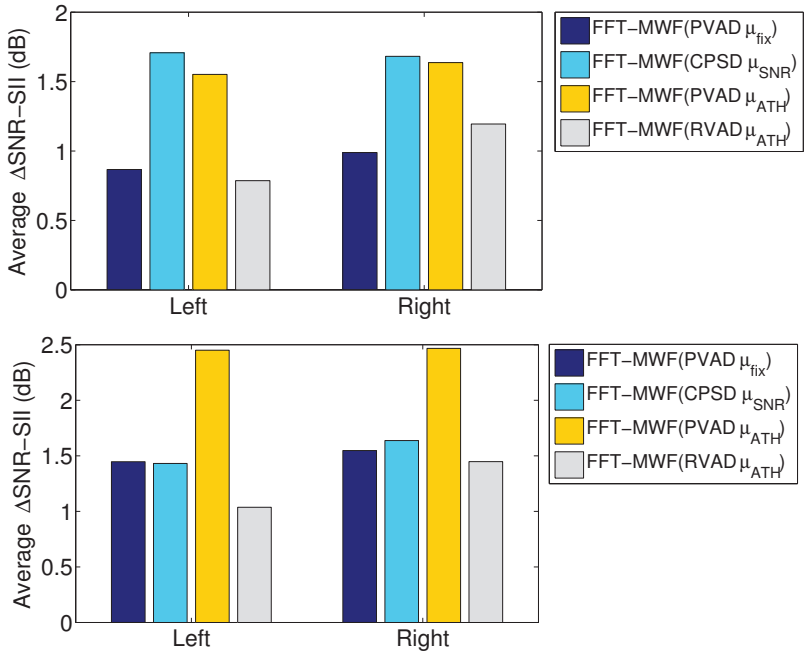

Figure 47: SNR improvement of FFT-MWF implemented with different strategies to estimate the statistics: VAD-based estimation and fixed μ (PVADμ_{fix}), CPSD-based estimation and adaptive μ (MWF-CPSDμ_{SNR}), and MWF framework based on auditory masking thresholds (MWF-μ_{ATH}). All implementations use an FFT length $L = 128$ and the SDW-MWF framework to compute the weights. Top: Getting in a cafeteria. Bottom: Getting out a cafeteria.

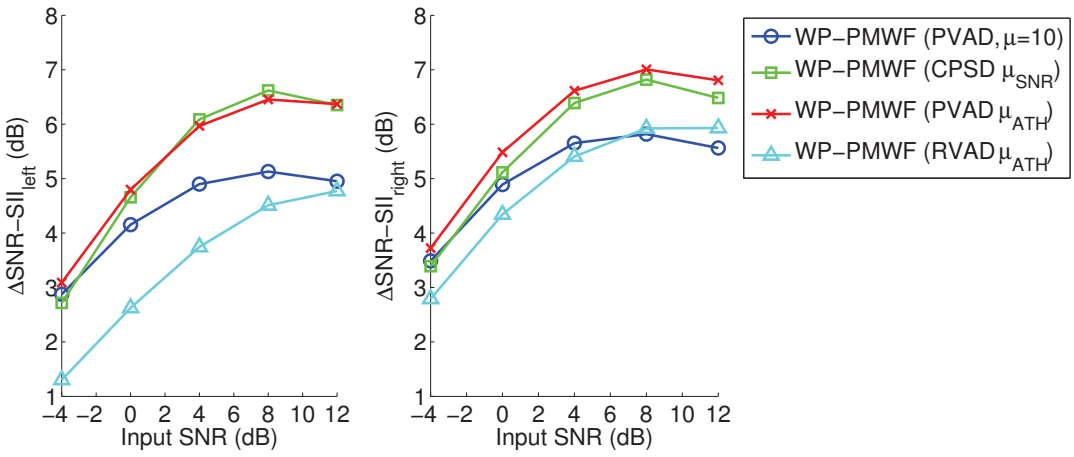

Figure 48: SNR improvement of WP-PMWF implemented with different strategies to estimate the statistics: VAD-based estimation and fixed μ (PVADμ_{fix}), CPSD-based estimation and adaptive μ (MWF-CPSDμ_{SNR}), and MWF framework based on auditory masking thresholds (MWF-μ_{ATH}). All MWF implementations use a WP with db8, and a frame length of $L = 128$. Scenario: Babble noise.

and MWF-CPSDμ_{SNR} provide similar performance. The above behaviors are also present for other metrics such as noise reduction (NPLR) (Figure 49). Thus, in PMWF, MWF-CPSDμ_{SNR} can replace MWF-μ_{ATH} but using less number of operations.

6.3.5 Performance Under Reverberant Conditions

The performance of any noise reduction algorithm is degraded when reverberation is present in the signal. So far, the WP-PMWF method provides good SNR improvement, good noise reduction, acceptable sound quality, and fewer operations than an FFT-based processing. All previous results were conducted under non-reverberant scenarios.

To analyze the effect of reverberation on the performance of the proposed method (WP-PMWF), simulations under 4 different reverberant rooms and babble noise are carried out. The WP-PMWF method is implemented using VAD-based statistics estimation and fixed $\mu = 10$. For comparison purposes, the performance of an FFT-based implementation using VAD-based statistics estimation, fixed $\mu = 5$, and frame length $L = 128$ is included. Results showed that the performance of the proposed method is degraded when the reverberation is increased (Figure 50) as expected. However, the degradation in the SNR improvement is not significant ($<$ 2dB), and the SNR improvement provided by WP-PMWF is still higher

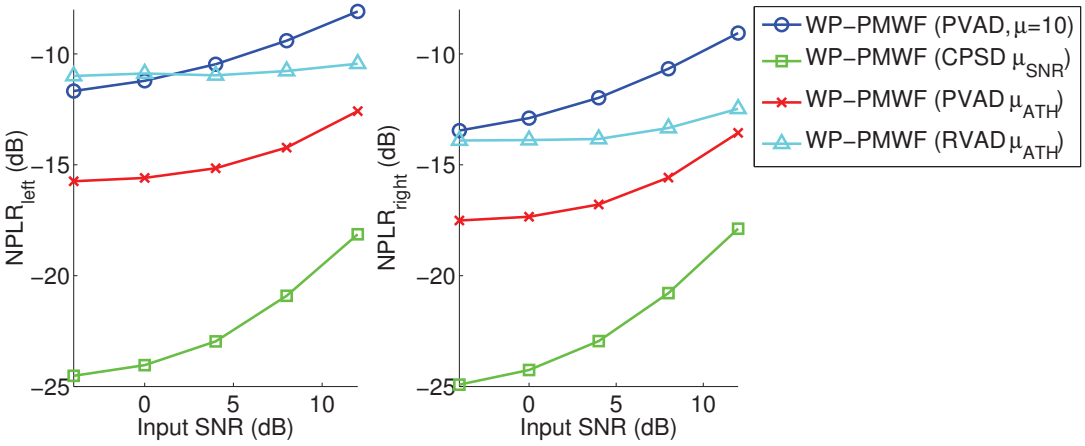

Figure 49: Noise reduction of WP-PMWF implemented with different strategies to estimate the statistics: VAD-based estimation and fixed μ (PVADμ_{fix}), CPSD-based estimation and adaptive μ (MWF-CPSDμ_{SNR}), and MWF framework based on auditory masking thresholds (MWF-μ_{ATH}). All MWF implementations use a WP with db8, and a frame length of $L = 128$. Scenario: Babble noise.

than that of the FFT-based processing. In terms of noise reduction (NPLR), there is more noise reduction for the room with higher reverberation (Figure 51). Hence, the proposed method provides an excellent performance under these environments.

6.3.6 MWF Framework Based on Binary Masking

Section 5.5 described a method to improve speech intelligibility in the MWF-based method. The proposed method, called MWF-IDBM, is the combination of a MWF method and an ideal binary mask (IDBM). To identify the most suitable way to generate the binaural binary mask, three strategies are tested: a) independent ideal binary mask for each channel; b) AND combination of independent ideal masks; c) OR combination of independent ideal masks. These methods are initially tested for an FFT-based processing to identify the feasibility of the proposed strategy and to avoid implementation issues due to the multirate processing in the WP-PMWF method. In particular, an FFT length of 512 samples is used to obtain high-resolution binary masks, and the ideal binary mask is created for a 0 dB SNR threshold taking the values $\{0, 1\}$. Results for the MWF-IDBM method using WP-based implementation are discussed later.

The performance of the base method (MWF), the ideal binary mask method (IDBM),

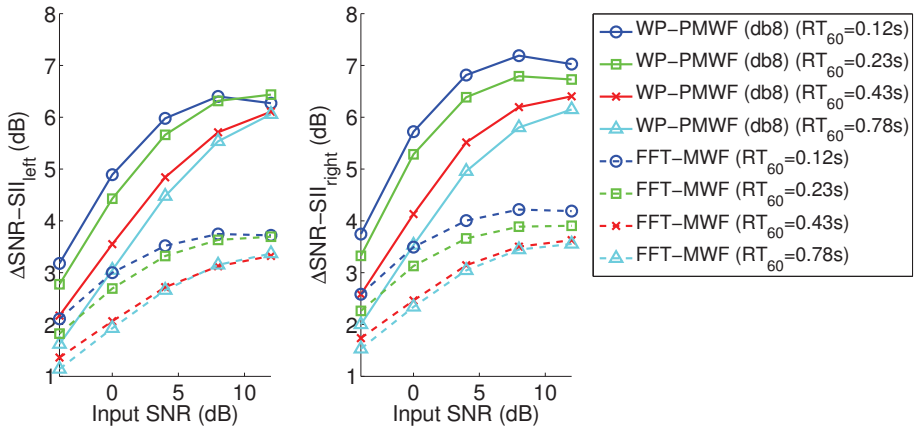

Figure 50: SNR improvement for WP-PMWF under 4 reverberant rooms.

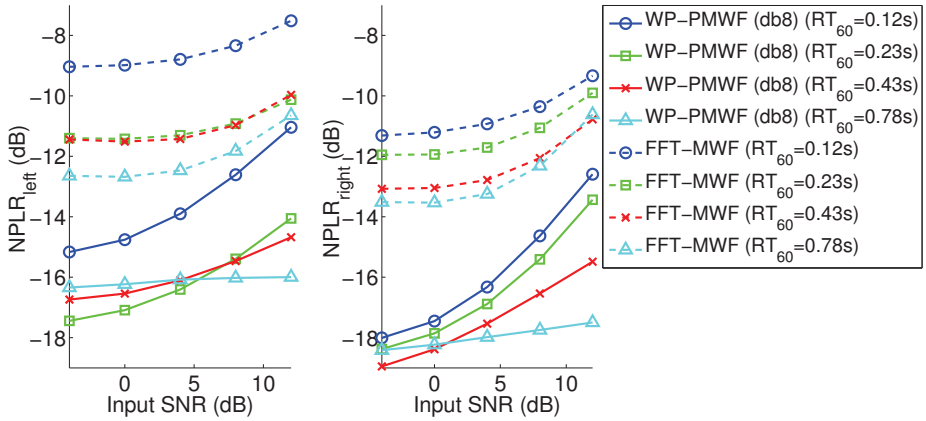

Figure 51: Noise reduction (NPLR) for WP-PMWF under 4 reverberant rooms.

and the proposed method (MWF+IDBM) using the three binaural binary-mask generation strategies described above is presented in Figures 52-55 for babble, small car, and traffic noise scenarios. The following conclusions are drawn from these figures:

- For all scenarios, the combined solution (MWF-IDBM) shows an improvement for all metrics compared to the base method (MWF). The improvement on the ΔSNR-SII and NPLR metrics is very significant, around 12 dB for ΔSNR-SII and 20 dB for NPLR. A similar benefit is found for the improvement in the speech intelligibility metric (I3). For example, under babble noise scenario at -5dB input SNR, the speech intelligibility in the base method, I3=0.30, is increased to I3=0.90 (Figure 52).

- The speech intelligibility improvement provided by the base method, MWF, is acceptable for all scenarios and SNR conditions, except for babble noise at -5dB input SNR (Figure 52). This suggests that the proposed method should be enabled only under highly-noise environments such as babble noise for input SNR $<$ 0dB.

- Since IDBM zeroes out the noise-only T-F bins, an improvement in the ΔSNR-SII and NPLR metrics is expected for this method. These metrics can be also improved with the proposed method. In IDBM, the T-F bins identified as speech-only by the ideal binary mask may contain noise, and this noise component cannot be reduced by IDBM. However, when IDBM is combined with MWF, the MWF gains apply additional noise reduction to these T-F bins, and then additional noise reduction is achieved. This fact is verified in the NPLR of the Figures 52-55. The above features show the potential benefit of the proposed solution with respect to a stand-alone IDBM solution.

- The AND-combined mask reduces the performance of the IDBM and MWF-IDBM methods significantly. This is explained by the removal of T-F bins identified as speech T-F bins in the better ear but irrelevant in the other ear.

- The performance using independent and OR-combined masks is similar. This result is particularly important for a binaural hearing aid since the OR-combined masks may involve the exchange of information through the wireless link. Therefore, independent masks are a preferable and sufficient strategy to accomplish the goals of SNR improvement, noise reduction, and speech intelligibility improvement.

The above results have been discussed for the MWF-IDBM method implemented with FFT processing. Further tests conducted in this research showed that the proposed strategy is also useful for the WP-PMWF framework. In this case, the ideal binary mask is created in the WP domain using a SNR threshold of $\theta_{th} = -2$ dB instead of 0 dB. Figures 56-57 show the performance of original WP-PMWF implementation, IDBM, and the combined solution WP-PMWF-IDBM. The WP-PMWF-IDBM implementation uses independent ideal binary

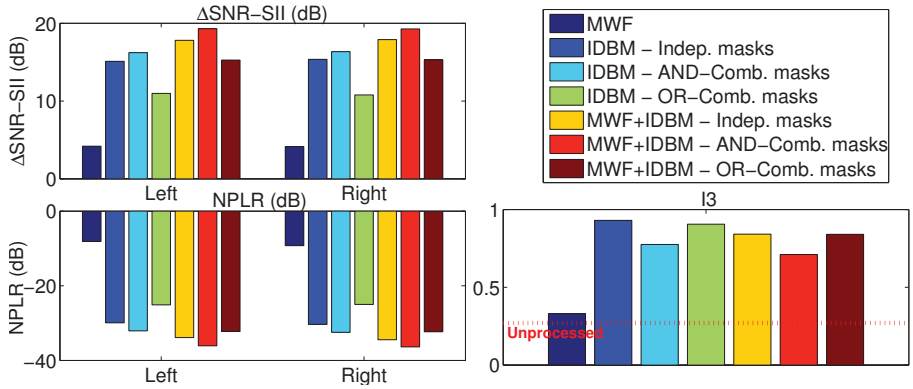

Figure 52: Performance of different binaural mask generation strategies in the MWF-IDBM method under babble noise scenario at -5 dB input SNR.

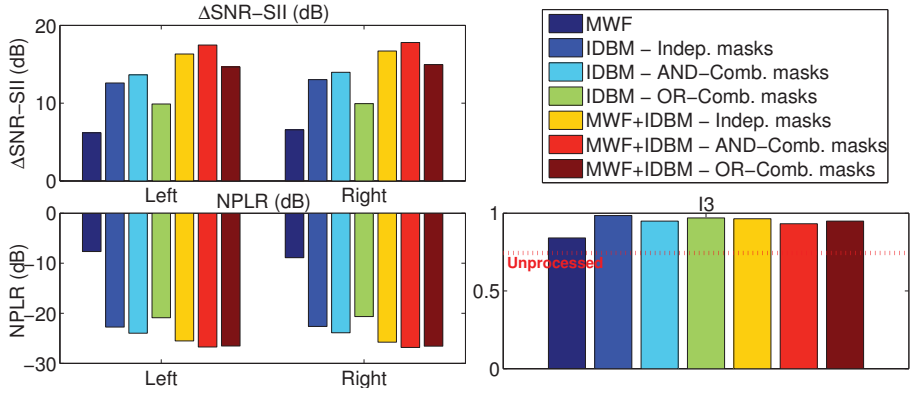

Figure 53: Performance of different binaural mask generation strategies in the MWF-IDBM method under babble noise scenario at 0 dB input SNR.

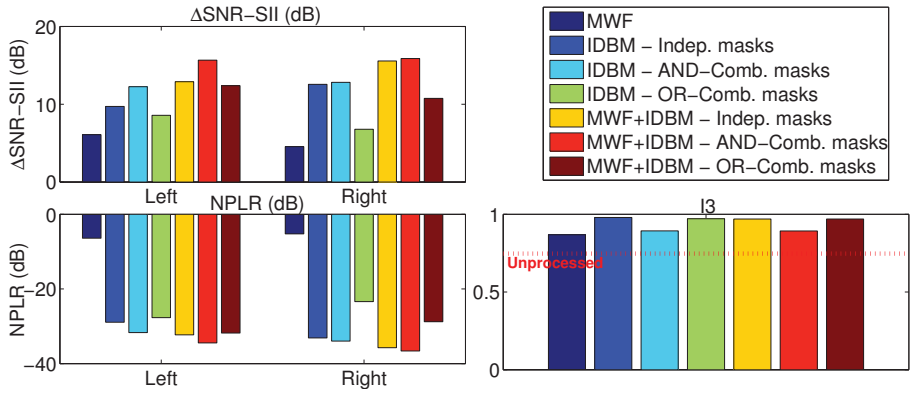

Figure 54: Performance of different binaural mask generation strategies in the MWF-IDBM method under small car noise scenario at -5 dB input SNR.

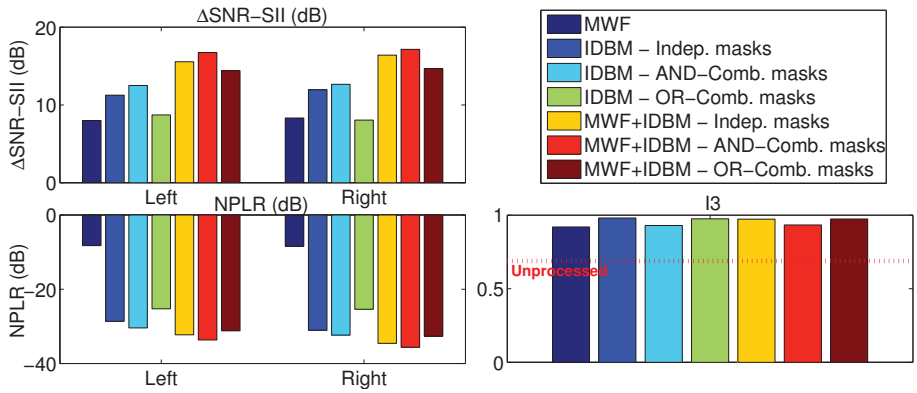

Figure 55: Performance of different binaural mask generation strategies in the MWF-IDBM method under traffic noise scenario at -5 dB input SNR.

masks for two ranges of the mask $\{0, 1\}$ and $\{0.2, 1\}$. The second range improves the sound quality as explained next.

For very low input SNR (e.g., -5 dB), audible musical artifacts are present in the enhanced signals by a stand-alone IDBM. Informal listening tests showed that these artifacts are also present in the proposed method when the ideal binary mask takes values $\{0, 1\}$. These musical artifacts are identified in the literature as the result of an over-subtraction effect. Hence, setting the minimum value of the mask to $\eta \neq 0$ may reduce these audible artifacts but degrade the algorithm performance (Figures 56-57). Although the performance of the proposed method in terms of SNR improvement and noise reduction for $\eta = 0.2$ is worse than for $\eta = 0$, this performance is significantly better than the performance of the original WP-PMWF method. In terms of the speech intelligibility, the I3 metric does not exhibit significant variation when $\eta = 0$ is replaced by $\eta = 0.2$.

A subjective test based on MUSHRA is conducted to assess the subjective sound quality. In this test, the subject is asked to grade the overall sound quality using a scale in the range 0-100. Six subjects participated in the experiment. The following algorithms were tested in the subjective test: MWF, IDBM with mask range $\{0, 1\}$, IDBM with mask range $\{0.2, 1\}$, MWF+IDBM with mask range $\{0, 1\}$, and MWF+IDBM with mask range $\{0.2, 1\}$. For babble noise scenario at -5 dB input SNR, IDBM and the MWF+IDBM are scored higher than the original MWF method (Figure 58a). This supports the claim that the proposed method can preserve the overall sound quality and at the same time improve

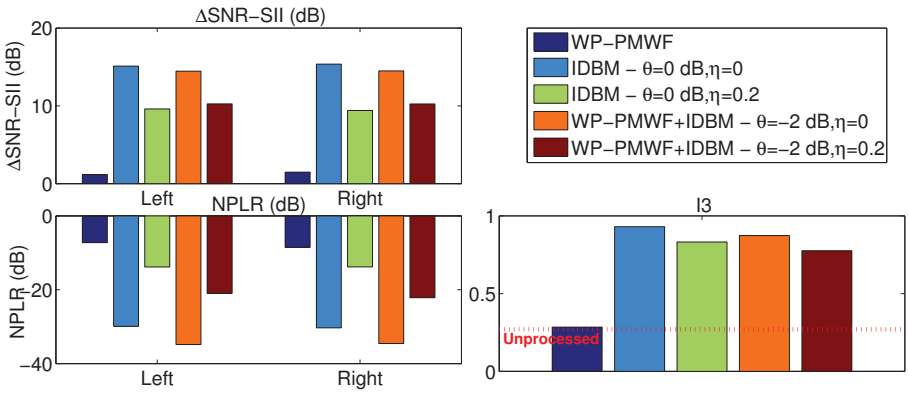

Figure 56: Performance of the WP-PMWF-IDBM method under babble noise scenario at -5 dB input SNR.

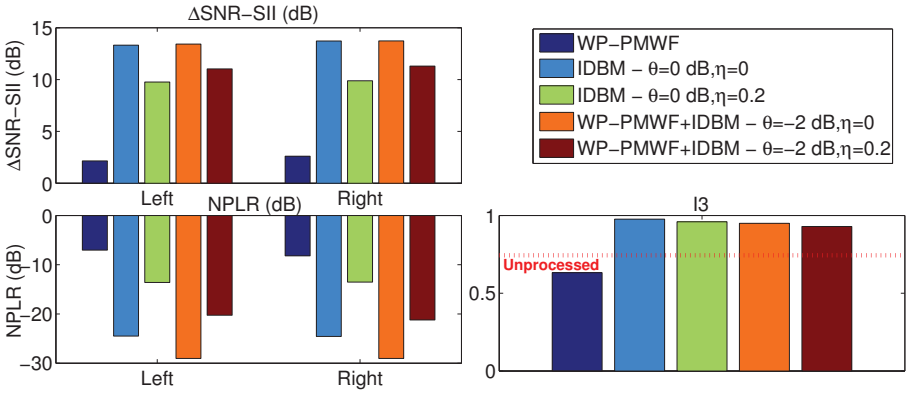

Figure 57: Performance of the WP-PMWF-IDBM method under babble noise scenario at 0 dB input SNR.

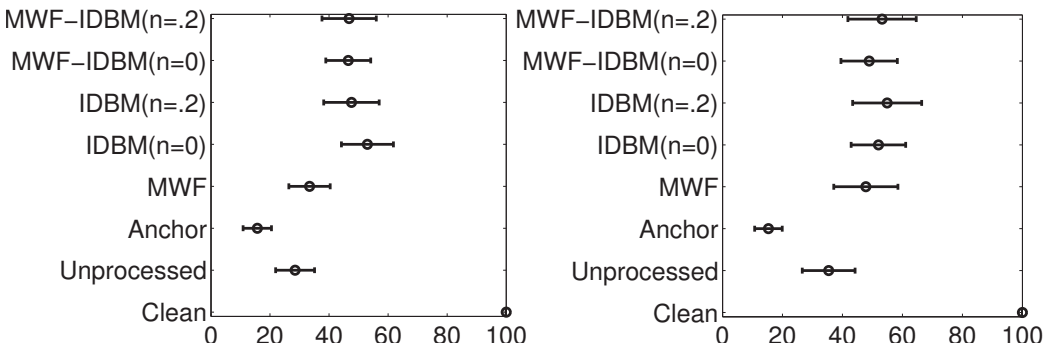

Figure 58: Overall subjective sound quality of the MWF, IDBM, and MWF-IDBM methods under babble noise scenario at two different input SNR: -5 dB (left) and 0 dB (right).

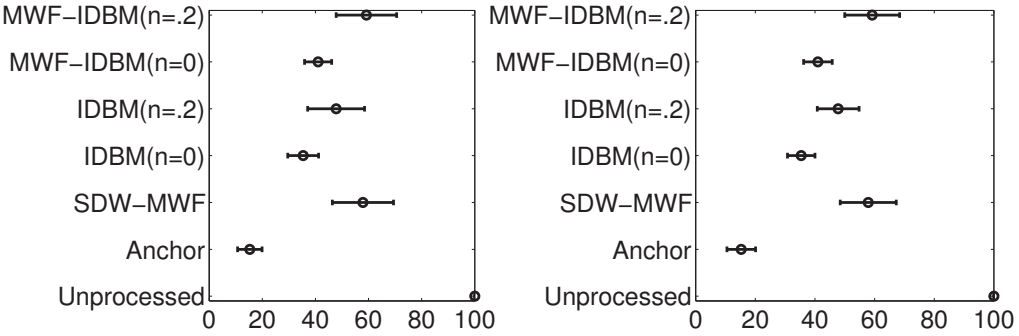

Figure 59: Background sound quality of the MWF, IDBM, and MWF-IDBM methods under babble noise scenario at two different input SNR: -5 dB (left) and 0 dB (right).

the speech intelligibility. For babble noise scenario at 0 dB input SNR (Figure 58a), the sound quality for all methods is rated similar. This result is consistent with the fact that musical artifacts in IDBM and MWF+IDBM for a mask range $\{0, 1\}$ are not as strong as for the -5 dB case, and the noise level offered by the original MWF is comfortable for the subject. Another subjective test was conducted to grade the quality of the background noise and to identify any perceptual difference between the IDBM and the MWF+IDBM methods (Figure 59). Results of this test support the claim that the usage of $\eta \neq 0$ provides better sound quality, and the proposed method outperforms an IDBM method in terms of sound quality. In conclusion, the proposed method using WP-based processing and binary mask is very promising.

Up to this point the binary mask has been generated using ideal information, i.e., having access to the clean and noisy signal. In a practical implementation, this binary mask is constructed from the noisy signal, which introduces mask estimation errors that degrade the performance of the proposed algorithm. Mask estimation errors are categorized in type I and type II errors. Type I errors, or false alarm, corresponds to 1's in the estimated mask that are 0's in the ideal mask. Type II errors, or miss, are 0's in the estimated mask that are 1's in the ideal mask. Simulations of the proposed method under different amounts of type I and type II errors are conducted to identify the accuracy required for the mask estimation. These simulations showed that speech intelligibility (I3 score) is degraded more significantly for type I errors than for type II errors. In particular, type I and type II errors must be kept below 15% and 40%, respectively, to achieve speech intelligibility greater than 50%. This result agrees with the study in [45] for monaural IDBM. Although ΔSNR-SII and NPLR metrics are also degraded by the introduction of mask estimation errors, this degradation is not as significant as for the I3 score.

This research explores three mask generation methods:

1. Mask generation method described by Roman *et al.* [80] (OIR-Mask). This method uses an adaptive filter to cancel out the target signal and so to obtain a noise estimate $\hat{v}$. The input and desired signals required by the adaptive filters are the signals at the left and right channels, y_L and y_R, respectively. Then, this noise estimate is used to compute the output-to-input energy ratio (OIR) for the left and right channel, which are defined by $OIR_L(k,l) = |\hat{v}(k,l)|^2 / |y_L(k,l)|^2$ and $OIR_R(k,l) = |\hat{v}(k,l)|^2 / |y_R(k,l)|^2$. Finally, a threshold is applied to the OIRs to obtain the mask at the left and right channel.

2. Mask generation based on the decision-directed *a priori* SNR estimation rule proposed by Ephraim and Malah (ASNR-Mask) [16]. In this case, the local SNR for the left channel is given by

$$SNR_L(k,l) = \alpha \frac{|z_{MWF-L}(k,l-1)|^2}{\boldsymbol{e}_L^H \boldsymbol{R}_v(k,l)\boldsymbol{e}_L} + (1-\alpha)\max\left[\frac{|y_L(k,l)|^2}{\boldsymbol{e}_L^H \boldsymbol{R}_v(k,l)\boldsymbol{e}_L} - 1, 0\right] \quad (44)$$

where $\boldsymbol{e}_L^H \boldsymbol{R}_v(k,l)\boldsymbol{e}_L$ is the noise power estimate computed from the noise correlation

matrix used by MWF method; z_{MWF-L} is the output of the MWF method at the left channel. A similar expression is obtained for the right channel. The masks are finally generated by setting a threshold to the local SNR.

3. Mask generation based on blind source separation (BSS-Mask). In this case, the BSS algorithm proposed in (2)-(5) (Section 4.2), is used to obtain a noise estimate $\hat{v}$, and this noise estimate is used to obtain the local SNR by $SNR_L(k,l) = \max\left\{|y_L(k,l)|^2 / |\hat{v}(k,l)|^2 - 1, 0\right\}$ and $SNR_R(k,l) = \max\left\{|y_R(k,l)|^2 / |\hat{v}(k,l)|^2 - 1, 0\right\}$. The masks are finally generated by setting a threshold to the local SNR.

The mask estimation method based on *a priori* SNR estimation (ASNR-Mask) is found to be impractical because type I errors are very high, and these errors must be kept below 15%. On the contrary, the BSS-Mask and OIR-Mask estimation methods provide low amount of type I errors (<12%) but high amount of type II errors (∼60%). In the OIR-Mask method, the estimation errors for the channel where the target signal is weaker are larger than in the BSS-Mask method. Post processing applied to the estimated masks is used to reduce type II errors by clustering isolated 1's. Two post processing strategies are explored in this research, median filter and mean filter. The mean filter is found to improve speech quality by reducing processing artifacts and producing more pleasant sounds. Figure 60 shows the performance obtained for the BSS-Mask and OIR-Mask methods. In terms of SNR improvement (ΔSNR-SII) and noise reduction (NPLR), the proposed method using estimated masks improves the performance of the original MWF method. However, in terms of speech intelligibility (I3), the speech intelligibility of the original MWF method is not improved by any of the proposed mask estimation methods. This result is a consequence of large amount of type II errors in the proposed mask estimation methods. The latter suggests further research to obtain a reliable mask estimation.

6.4 Performance of Transmission Bandwidth Reduction in WP-PMWF

The WP-PMWF strategy provides good SNR improvement and noise reduction, and involves less computational complexity than the BSS-PP strategy or other MWF strategies.

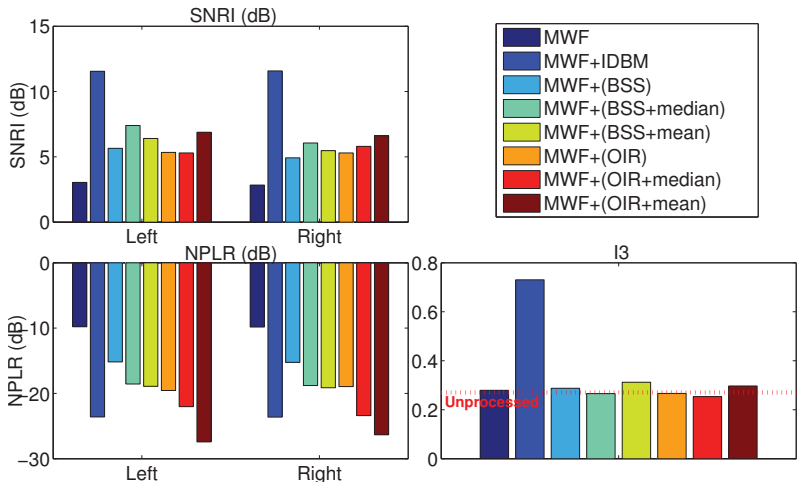

Figure 60: Performance of the proposed MWF-IDBM method using mask estimation based on output-to-input energy ratio (OIR-Mask) and blind source separation (BSS-Mask) under babble noise at -5 dB input SNR.

The use of multirate processing in WP-PMWF gives an opportunity to obtain a reduced-bandwidth WP-PMWF implementation. This solution was analyzed in the Section 5.2.2, in which only the low-frequency sub-bands are transmitted to the contralateral hearing aid.

To explore the impact of transmitting different number of sub-bands (S) and channels (T) in WP-PMWF, simulations for different configurations of T and S are conducted under babble noise scenario. All experiments use $M = 2$ microphones per hearing aid, and the sampling rate is 22 kHz. For this sampling rate, the total number of sub-bands is 24. Figure 61 presents the SNR improvement and objective quality for different configurations. These results show no significant degradation in the performance (SNR improvement and sound quality) when only one channel (T=1) is transmitted to the contralateral hearing aid. This result is reported previously by [98], in which the transmission of a single channel provides similar performance to the method employing full transmission of channels. In addition, there is a small performance reduction when the number of transmitted sub-bands is reduced from S=24 (full transmission) to S=6 (transmission of the sub-bands for frequencies below 1.5 kHz). These results suggest that the proposed bandwidth reduction in WP-PMWF is very promising.

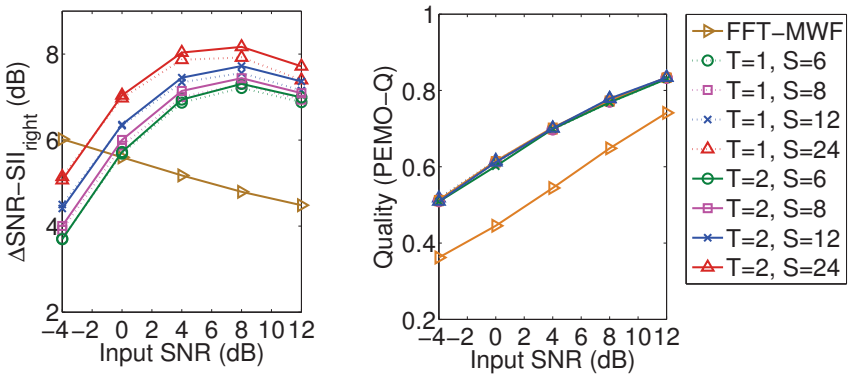

Figure 61: SNR improvement (left) and objective quality (right) for different number of transmitted sub-bands (S) and channels (T) in WP-PMWF. WP-PMWF uses db8 and $f_s = 22$ kHz. Plots for $S = 24$ corresponds to the transmission of all sub-bands.

Table 4: Bandwidth reduction in WP-PMWF for different transmitted sub-bands (S) and channels (T). NTxSamp: number of samples transmitted for each input sample. Bandwidth estimated for 16-bit samples without encoding at 22 kHz sampling rate.

T	S	NTxSamp	Bandwidth (bps)	%Reduction
1	6	0.0469	16509	95%
1	12	0.2188	77018	78%
1	24	1.0000	352000	0%
2	6	0.0938	33018	91%
2	12	0.4375	154000	56%
2	24	2.0000	704000	0%

The proposed bandwidth reduction provides an impressive reduction compared to the full rate transmission. Assuming that every transmitted sample is 16 bits without coding, the number of bits per second of the wireless link in one direction for different configurations is presented in the Table 4. This bandwidth reduction also reduces slightly the computational cost (Fig. 62) compared to the original WP-PMWF method. The computational cost for the FFT-based implementation is also included for comparison purposes. This plot suggests that the proposed method is a good candidate to reduce both transmission bandwidth and computational cost.

6.5 Comparison of the Proposed BSS and MWF Based Methods

This research introduced two binaural noise-reduction methods inspired in perceptual processing: BSS-PP and PMWF. The performance of these methods is presented independently in the Sections 6.2 and 6.3, where both proposed methods outperform existing binaural noise

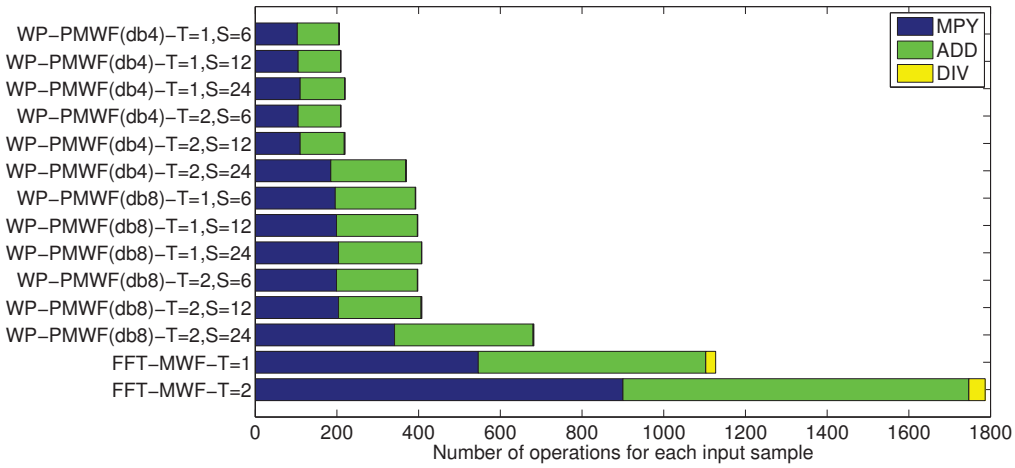

Figure 62: Number of operations for each input sample at different number of transmitted sub-bands (S) and channels (T) in WP-PMWF. WP-PMWF is implemented with a WP using db8. All implementations assume $M = 2$ microphones per hearing aid. Sampling frequency $f_s = 22$kHz. Plots for $S = 24$ corresponds to the transmission of all sub-bands.

reduction methods. In terms of computational complexity, the BSS-PP and FW-PMWF methods have a computational complexity similar to the FFT-based MWF implementation (Figures 10 and 17). On the contrary, WP-PMWF provides a significant complexity reduction (Fig. 17) as well as significant reduction in the transmission bandwidth.

A comparison of the BSS and MWF based methods under babble noise scenario is shown in the Figures 63-65. In terms of SNR improvement (Fig. 63) and noise reduction (Fig. 64), the BSS-PP method provides better performance. However, BSS-PP introduces degradation in the sound quality with respect to PMWF (Fig. 65). This fact is previously identified through the subjective test on the Section 6.2, in which BSS-PP is compared with respect to other binaural BSS-based noise-reduction methods and the MWF-N method.

Among the two PMWF methods, the implementation using frequency-warped filters provides slightly better performance than the wavelet-packet-based implementation in terms of SNR improvement and objective sound quality. The latter is verified through subjective quality assessments (Fig. 66) using the MUSHRA protocol [26] to grade the overall sound quality. Although the subjective test shows that FW-PMWF provides better sound quality followed by WP-PMWF, the WP-based implementation is a preferable implementation due to the reduction in both computational complexity and transmission bandwidth.

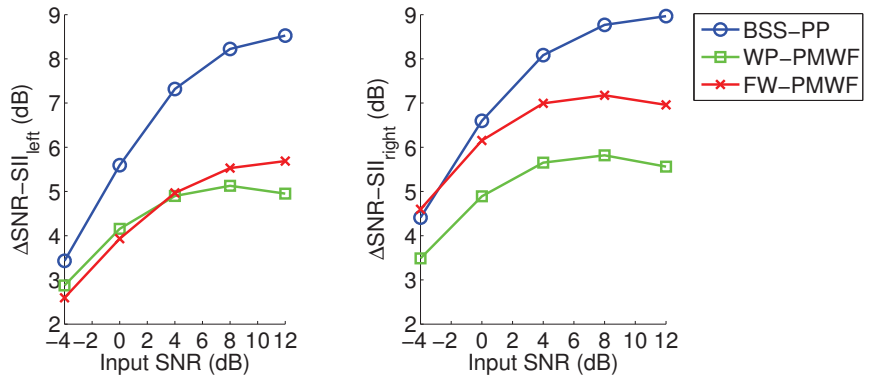

Figure 63: SNR improvement for the proposed methods, BSS-PP and PMWF, under babble noise scenario.

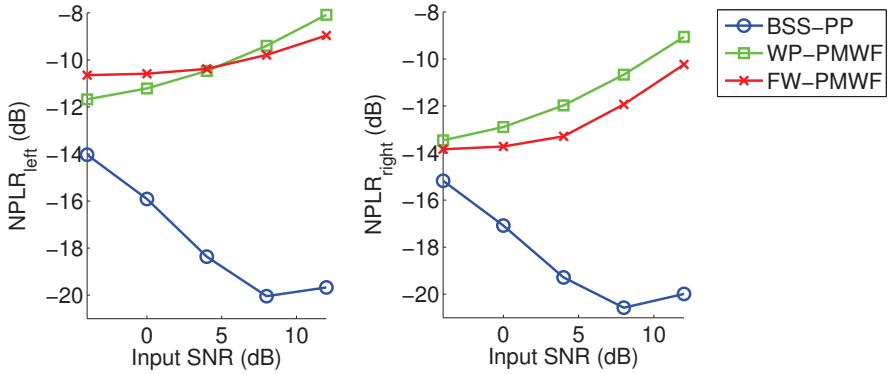

Figure 64: Noise reduction for the proposed methods, BSS-PP and PMWF, under babble noise scenario.

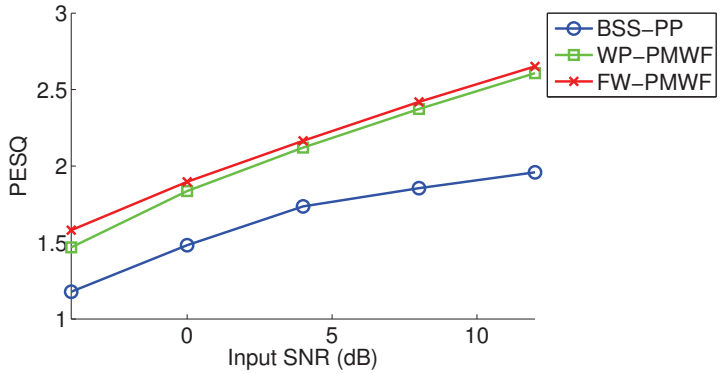

Figure 65: Objective quality for the proposed methods, BSS-PP and PMWF, under babble noise scenario.

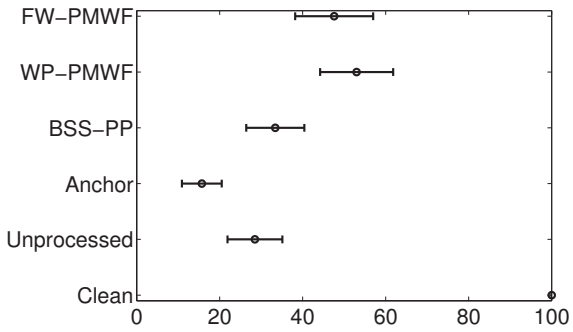

Figure 66: Subjective test for the proposed methods: BSS-PP and PMWF.

6.6 Summary

In this chapter, the binaural noise-reduction methods based on BSS and MWF, proposed in the Chapters 4 and 5, are analyzed under different scenarios including reverberant and non-reverberant conditions. Among the proposed methods, one of the most successful strategies is the replacement of the FFT processing by an auditory filterbank implemented by wavelet packets (WP-PMWF). A summary of the performance of the proposed methods is presented in the Table 5. The success of the WP-PMWF method relies on the good performance in terms of noise reduction and sound quality as well as the low computational complexity and transmission bandwidth. This performance is the result of higher low-frequency resolution in the PMWF method compared to the FFT-based MWF. Although the SNR improvement at very low input SNR provided by WP-PMWF is comparable to the SNR improvement of FFT-based MWF, the noise reduction in WP-PMWF at these input SNR conditions is significantly better. The other PMWF methods and the BSS-PP method provide good noise reduction and acceptable sound quality, but only WP-PMWF provides small computational complexity and reduction of the transmission bandwidth.

The effect of different WP-PMWF parameters is also analyzed, concluding that small frame length, L, is required to achieve better noise reduction and sound quality. In addition, the order of the mother wavelet must be high to achieve good sound quality. It is recommended to use a mother wavelet Daubechies for an order $n \geq 4$. Other wavelet families such as Symlets and Coiflets can be used, obtaining similar sound quality but involving higher computational cost.

Table 5: Comparison between the proposed methods: BSS-PP and PMWF (wavelet packet–WP and frequency-warped filters–FW implementations)

	BSS-PP	**WP-PMWF**	**FW-PMWF**
Noise reduction	Excellent	Good	Good
Speech quality	Fair	Good	Good
Preservation of localization cues	Yes	Yes	Yes
Computational complexity	Similar to FFT-based MWF	Lower than FFT-based MWF	Slightly above FFT-based MWF
Latency (for $f_s = 16$ kHz)	One sample (0.0625ms)	Depends on the sub-band. 6ms for low-freq. sub-bands and 0.4ms for high-freq. sub-bands	One sample (0.0625ms)
Transmission-bandwidth reduction	Not possible in the current solution	Possible	Not possible in the current solution

Another MWF processing strategy presented in this chapter is the estimation of the second-order statistics using a noise cross-PSD estimator (CPSD) and an adaptive trade-off parameter μ based on the frame SNR. This strategy, called MWF-CPSDμ_{SNR}, does not require a VAD, and it is shown to provide significant benefits over a VAD-based statistics estimation, particularly under highly non-stationary environments. This strategy is tested in FFT-based MWF and PMWF. In both cases, the MWF-CPSDμ_{SNR} implementation improves the performance of a VAD-based implementation.

The computation of μ in the MWF-CPSDμ_{SNR} strategy has a close relationship with a MWF framework based on auditory masking thresholds (MWF-μ_{ATH}), but the expression in MWF-CPSDμ_{SNR} involves less computational cost. Simulations showed the strong influence of the estimation errors on the performance of MWF-μ_{ATH}. Besides, the upper-bound performance of MWF-μ_{ATH} can be reached by MWF-CPSDμ_{SNR}. Thus, MWF-CPSDμ_{SNR} is a promising method to implement any MWF technique, and it can replace a MWF framework based on auditory masking thresholds.

Finally, the method proposed to improve speech intelligibility (MWF-IDBM) shows to be useful only for some scenarios, such as babble noise at very low input SNR (< 0 dB).

For these scenarios, MWF-IDBM provides an excellent performance under ideal conditions. However, under real estimation of the binary mask, the speech intelligibility is dramatically reduced. Hence, further research is required in the development of practical methods for binaural mask generation.

Chapter VII

PRACTICAL IMPLEMENTATION OF MWF

Section 2.3 discussed different implementation challenges for a binaural noise-reduction methods in a digital hearing aid: computational cost, latency, processing artifacts, estimation of parameters, power consumption, and transmission bandwidth. Most of these issues were addressed in the design of the methods proposed in the Chapters 4 and 5. In this chapter, additional solutions to reduce computational cost and processing artifacts are discussed.

The two techniques proposed in this research, BSS-PP and PMWF, do not employ FFT processing. Therefore, the audible artifacts inherent to the FFT convolution are absent in the two proposed techniques. However, there is another source of artifacts coming from the processing. In BSS-PP, audible artifacts may come from the auditory filterbank (implemented with forth-order IIR filters). Since the filterbank specification meets the critical band criteria, audible artifacts are minimized with this processing. On the other hand, WP-PMWF uses wavelet packet for analysis and synthesis. The wavelet packet is known to be a perfect reconstruction architecture. Therefore, WP-PMWF is also a processing free of audible artifacts. To ensure the absence of processing artifacts in the reduction of computational complexity and transmission bandwidth in PMWF, this chapter introduces additional methods that employ perfect reconstruction processing.

An analysis of the computational cost for the proposed approaches showed that the PMWF implementation using wavelet packets provides the smallest number of operations (Figure 17 in Section 5.6). To implement WP-PMWF in a DSP, more than 75% of the CPU resources are dedicated to the computation of the wavelet packet (Figure 18 in Section 5.6). Thus, hardware acceleration for the WP is an excellent strategy to reduce the CPU utilization. The majority of the available hardware acceleration for WP is commonly targeted for a specific WP tree, not necessary the WP tree used in this research. Thus, if

hardware acceleration in not an option, alternative ways to reduce the CPU utilization in the WP computation must be explored. One of these alternatives is analyzed in the Section 7.1, in which the WP is replaced by a discrete wavelet transform (DWT).

Although commercial DSP architectures does not include hardware acceleration for the WP or DWT, some DSP architectures include hardware acceleration for the FFT. Therefore, an FFT-based MWF implementation would be preferable for these architectures. Contrary to the WP-PMWF method, the FFT-based implementation involves more CPU utilization in the update of statistics and weight computation (Figure 18 in Section 5.6). A similar situation is present for the FW-PMWF method. Although FW-PMWF provides better performance than the FFT-based MWF implementation, its computational cost is higher than a FFT-based MWF implementation. Hence, a method to reduce the computational overhead due to the update of statistics and weight computation is presented in the Section 7.2. The proposed method can be used for both FFT-based MWF and FW-PMWF implementations.

The implementation of a wide range of existing binaural noise-reduction methods, including MWF, employ FFT-based block processing. Therefore, these methods can be easily implemented in DSP architectures that include FFT hardware acceleration. However, an FFT-based block processing introduces audible artifacts. These audible artifacts are the consequence of non-linear algorithms to generate the frequency-domain filter weights, which invalidates the condition to avoid circular convolution. We performed a detailed mathematical analysis of the artifacts, and propose two FFT-based block processing strategies to avoid audible artifacts, which are described in the Section 7.3.

7.1 Simplification of the Analysis/Synthesis Stage in PMWF

As shown in Fig. 18, the majority of the processing involved in WP-PMWF is related to the analysis and synthesis using a WP tree that resembles the auditory filterbank. The number of operations in WP-PMWF depends strongly on the mother wavelet. Different analysis conducted in Section 6.3 showed that the order of the mother wavelet must be $\geq$4 (e.g., db4) to get an acceptable sound quality. Hence, to reduce the number of operations

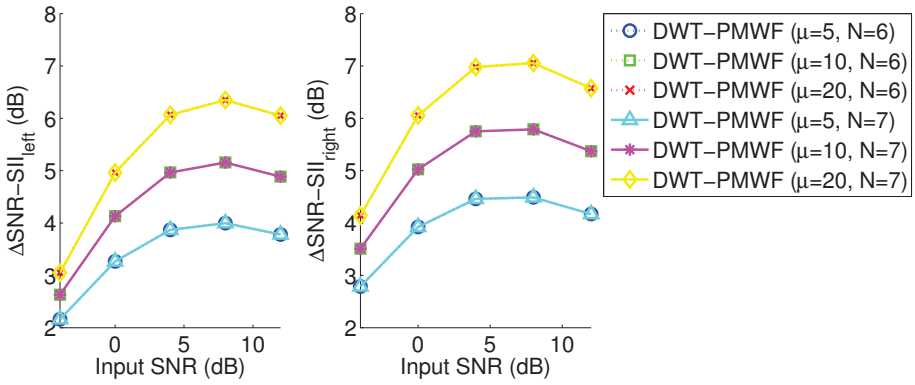

Figure 67: SNR improvement for DWT-PMWF under babble noise scenario using different number of decomposition levels and trade-off parameter μ.

in WP-PMWF, it is necessary to modify the WP tree. The most simple solution is to replace the WP tree by a DWT tree. This solution is called DWT-PMWF. In the DWT tree only the low frequency sub-bands are split using low-pass and high-pass filters, followed by down-samplers. Although the filterbank associated to the DWT tree does not resemble exactly an auditory filterbank, it provides high frequency resolution for the low-frequency sub-bands as in the auditory filterbank. In this sense, the performance of DWT-PMWF is expected to be similar to WP-PMWF but using less number of operations.

Simulations for DWT-PMWF are conducted for two number of decomposition levels, 6 and 7, i.e., 7 and 8 sub-bands. Results show no difference for the performance under babble noise in terms of SNR improvement and noise power level reduction (Figures 67 and 68). This result is a consequence of the decomposition tree used by the DWT. Increasing the number of decomposition levels in the DWT increases the number of low-frequency sub-bands. But these low-frequency sub-bands are related to very low frequencies, and no significant improvement is achieved by including a large number of low-frequency sub-bands. Likewise, compared to WP-PMWF, the DWT processing does not degraded significantly the performance of the WP-based PMWF implementation in terms of SNR improvement and noise power level reduction (Figures 69 and 70). However, the DWT computation involves one half the number of operations of the WP computation (Table 6). Hence, DWT-PMWF is a promising solution to achieve reduction in the computational complexity and to maintain the same performance of WP-PMWF.

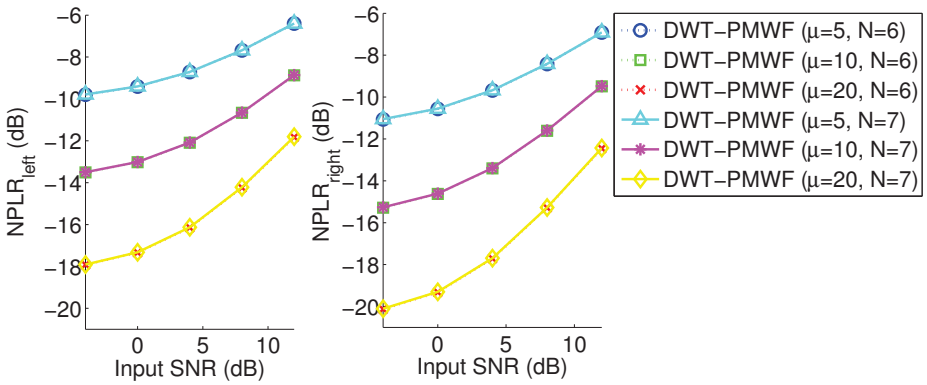

Figure 68: Noise reduction for DWT-PMWF under babble noise scenario using different number of decomposition levels and trade-off parameter μ.

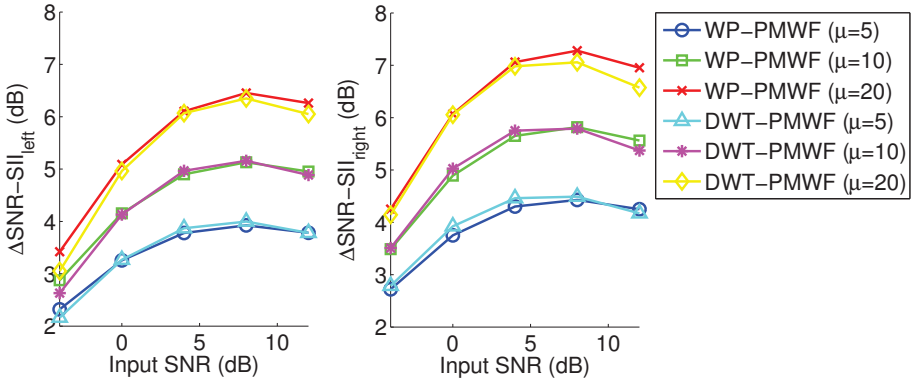

Figure 69: SNR improvement for DWT-PMWF and WP-PMWF under babble noise scenario for different trade-off parameters μ. DWT uses 6 decomposition levels.

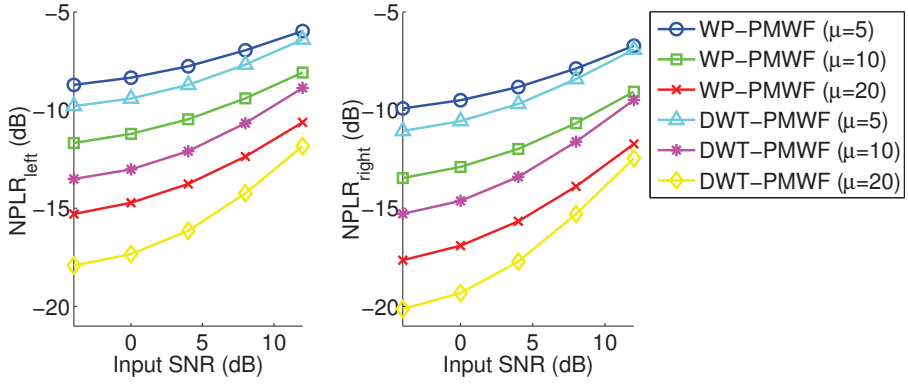

Figure 70: Noise reduction for DWT-PMWF and WP-PMWF under babble noise scenario for different trade-off parameters μ. DWT uses 6 decomposition levels.

Table 6: Number of MACs in the WP-based or DWT-based PMWF implementations is given by $TF \times Q \times L$, where Q is the filter length depending on the mother wavelet, L is the frame length, and TF is the factor related to the wavelet tree. The value of TF for both WP-based and DWT-based PMWF is shown below.

	TF
WP ($f_s = 16$ kHz)	$125/32 = 3.9$
WP ($f_s = 22$ kHz)	$317/64 = 4.59$
DWT (6 levels)	$63/32 = 1.97$
DWT (7 levels)	$127/64 = 1.98$

7.2 Recursive-Update MWF (RECUP-MWF)

WP-PMWF is introduced to improve the performance of an FFT-based MWF processing and to reduce the computational complexity, replacing the FFT by a wavelet packet. Some hearing-aid DSP architectures may include hardware acceleration for the FFT computation. Hence, the implementation of WP-PMWF may not be a good choice for these particular architectures. A simple way to reduce processing in the FFT-based processing is to reduce the FFT length. Section 6.3 shows that an FFT-based MWF solution using small frame length ($L = 32$) is not a good solution. In FFT-based processing nearly 75% of the CPU usage is related to statistics update and weight computation (Figure 18). The same situation is present in the FW-WPMF method. Therefore, to reduce computational complexity in these methods, it is necessary to explore alternative methods to reduce the overhead related to these functional stages of the algorithm.

7.2.1 Background

As shown in Section 5.1, weight computation involves solving a linear system of equations. It is widely known that Cholesky, LDU, or QR decompositions are efficient ways to solve a linear system of equations. In this research, all previous reports on computational complexity assumed the usage of Cholesky decomposition since the correlation matrices are Hermitian and positive definite. In the literature, there are reports on efficient algorithms to implement MWF. These algorithms are based on subspace [12, 91, 36], QR decomposition (QRD) [82, 81, 37], and steepest-descendent algorithms [92, 93]. A brief overview of the existing methods, and their advantages and disadvantages are described next.

In the subspace-based methods, generalized singular value decomposition (GSVD) is employed to obtain a decomposition in which the principal component is associated to the target speech while the other components are associated to the noise. A recursive GSVD method was initially proposed by Doclo and Moonen in [12]. A simplification of this method, called rank-one MWF (R1-MWF), was introduced in [91], and improved in [36] to reduce the estimation error on the principal component. All the above subspace methods require *a priori* information about the direction of arrival (DoA) of the target signal to identify the principal component. This DoA is assumed usually in the front, which limits one of the main advantages of MWF, the enhancement of the target signal coming from any arbitrary direction of arrival.

In [82], Rombouts and Moonen proposed a QRD-based MWF method. The method is stable and computationally more efficient than a SVD-based MWF. It can also be implemented efficiently in hardware using systolic arrays [81]. Although this method requires a VAD, Kim and Cho [37] proposed modifications to avoid the usage of a VAD. The main disadvantage of these methods is the implementation of SDW-MWF assuming μ fixed to 1. Another μ values cannot be included in those architectures. This dissertation shows that using an adaptive μ provides significant performance improvement (Sections 6.3.3, 6.3.4, and 6.3).

Spriet *et al.* proposed different methods based on the steepest-descendent algorithm [92, 93]. These algorithms perform a recursive update of the filter coefficients by means of a step related to the gradient of a cost function. Different from the subspace and QRD methods, these methods allow the usage o trade-off parameter μ. However, the method has been designed to enhance the target signal coming exclusively from the front. Although the algorithm may be modified to enhance the target signal coming from any arbitrary direction of arrival, this modification requires a DoA algorithm, which limits the performance of the algorithm. In addition, the method assumes that the environment changes slowly, which may not be the case of highly non-stationary scenarios such as babble noise.

The method proposed in this section is a solution that reduces computational complexity, allows an adaptive μ, enhances the target signal coming from any arbitrary DoA, and is

//Initialization

$\boldsymbol{R}^{-1}(f,0) = \boldsymbol{I}$

//Processing

$\boldsymbol{q}(f,l) = \boldsymbol{R}^{-1}(f,l-1)\boldsymbol{y}(f,l)$

if (voiced segment)

$\boldsymbol{r}_{x_L}(f,l) = \lambda \boldsymbol{r}_{x_L}(f,l-1) + (1-\lambda) y_L^*(f,l)\boldsymbol{y}(f,l)$

$\boldsymbol{r}_{x_R}(f,l) = \lambda \boldsymbol{r}_{x_R}(f,l-1) + (1-\lambda) y_R^*(f,l)\boldsymbol{y}(f,l)$

$\mu' = 1-\lambda$

else //unvoiced segment

$\boldsymbol{r}_{x_L}(f,l) = \lambda \boldsymbol{r}_{x_L}(f,l-1)$

$\boldsymbol{r}_{x_R}(f,l) = \lambda \boldsymbol{r}_{x_R}(f,l-1)$

$\mu' = \mu(1-\lambda)$

end

$\beta(f,l) = \frac{1}{\alpha + \mu' \boldsymbol{y}^H(f,l)\boldsymbol{q}(f,l)}$

$\boldsymbol{R}(f,l)^{-1} = \frac{1}{\alpha}\left[\boldsymbol{R}^{-1}(f,l-1) - \mu'\beta(f,l)\boldsymbol{q}(f,l)\boldsymbol{q}^H(f,l)\right]$

$z_L(f,l) = \beta(f,l)\boldsymbol{r}_{x_L}^H(f,l)\boldsymbol{q}(f,l)$

$z_R(f,l) = \beta(f,l)\boldsymbol{r}_{x_R}^H(f,l)\boldsymbol{q}(f,l)$

Figure 71: Recursive-Update MWF.

suitable for highly non-stationary environments. Figure 71 includes a summary of the proposed algorithm, which is derived from a recursive update of the inverse correlation matrix (Appendix C). The proposed method is called recursive-update MWF or RECUP-MWF.

7.2.2 Performance of RECUP-MWF

The main purpose of RECUP-MWF is the reduction of computational complexity in the weight computation of the SDW-MWF framework. The algorithm performance is very sensitivity to the forgetting factor α. This parameter comes from different approximations, and it is assumed to be close to 1 for the derivation of the algorithm. A mathematical analysis about the effect of this parameter on the stability of the algorithm shows that the value of α must be close to 1 [58]. A value of $\alpha = 0.995$ is chosen for all experiments.

Table 7 shows the number of operations involved in the RECUP-MWF implementation (Figure 80). For comparison purposes, the number of operations involved in the statistics update and weight computation, using a linear solver based on Cholesky decomposition, are also included in this table. The reduction in the number of multiplications and additions is significant for RECUP-MWF only when the number of microphones per hearing aid is $M \geq$

Table 7: Comparison between the number of multiplications (MPY), additions (ADD), and divisions (DIV) of the MWF implementation using Cholesky decomposition (Linear Solvers), and the RECUP-MWF implementation. The number of operations is presented for a different number of microphones, M, per hearing aid.

	MPY		ADD		DIV	
M	Lin. Solv.	RECUP-MWF	Lin. Solv.	RECUP-MWF	Lin. Solv.	RECUP-MWF
1	21	21	9	11	3	1
2	94	54	42	33	10	1
3	235	103	107	67	21	1

2. In addition, regardless M, the number of divisions is always one in the RECUP-MWF implementation, which is a significant reduction compared to the MWF implementation using Cholesky decomposition. Since divisions are time-consuming operations for a fixed-point DSP, the proposed method is a promising MWF implementation for these DSPs. The parameter β in the RECUP-MWF algorithm exhibits a variation in the range $\beta = [0, 1]$, with values nearly to 1. Thus, it is possible to get a division-free algorithm by using a look-up table to compute β or setting β to a fixed value, e.g., $\beta = 1$. When β is forced to 1, additional modifications are required to ensure a stable algorithm to update the inverse correlation matrix $\boldsymbol{R}^{-1}$. In particular, the diagonal elements of $\boldsymbol{R}^{-1}$ under this simplification become unstable. Therefore, the diagonal elements in the simplified version are assumed to be fixed values and no update rule is applied for these elements. On the contrary, the off-diagonal elements of $\boldsymbol{R}^{-1}$ are updated following the rules stated in (60, Appendix C). Although the simplified RECUP-MWF algorithm is a division-free algorithm, the number of additions and multiplications is reduced slightly compared to the Full RECUP-MWF implementation, and the sound quality is degraded.

The computational complexity for different MWF implementations and number of microphones per hearing aid (M) is presented in the Fig. 72. Four kinds of MWF architectures are reported in this figure: a) FFT-based implementations assuming an FFT length $L = 128$; b) PMWF implementations using wavelet packet (WP) assuming mother wavelet db4; c) PMWF implementations using frequency-warped (FW) filters for $N = 16$ all-pass filters in each filter chain. These architectures are reported for two different methods to compute the weights: linear solvers based on Cholesky decomposition, and RECUP-MWF.

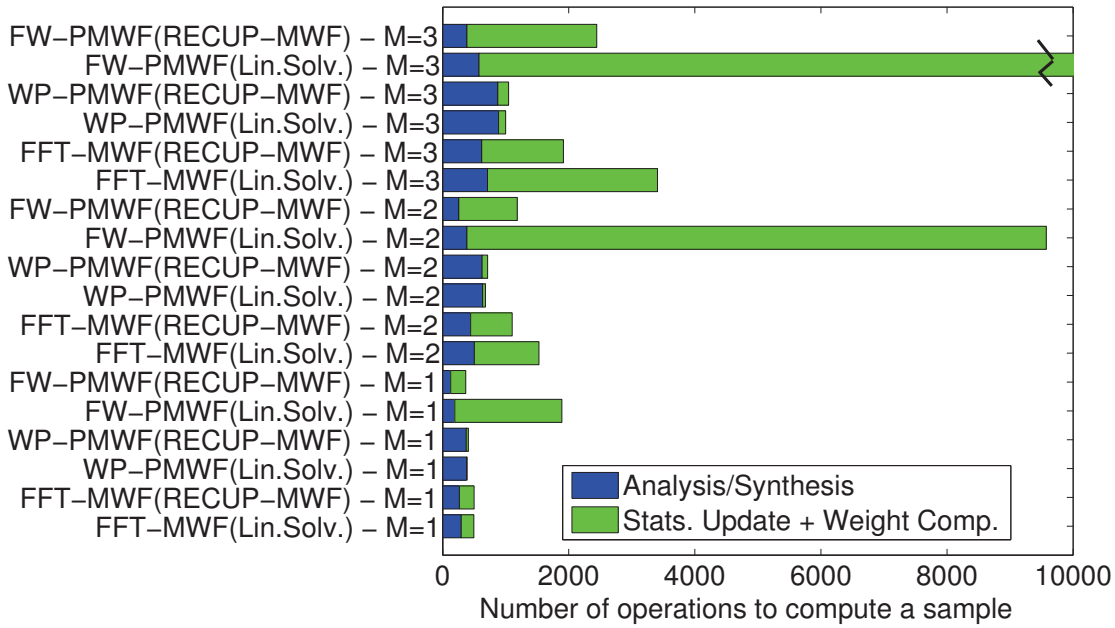

Figure 72: Computational cost for different MWF implementations and number of microphones per hearing aid (M). The cost is reported for three MWF architectures: FFT-based MWF, WP-PWMF, and FW-PMWF. These architectures are also reported for two methods to compute the weights: linear solvers based on Cholesky decomposition (Lin.Solv.), and RECUP-MWF. The number of operations of FW-PMWF using Cholesky decomposition for $M = 3$ exceeds the range of plotting.

As mentioned before, RECUP-MWF is not suitable for $M = 1$ but there is a significant complexity reduction when $M \geq 2$. RECUP-MWF provides a reduction around 25% in the FFT-based implementation for $M = 2$ microphones using linear solvers based on Cholesky decomposition, and this reduction is better for a larger number of microphones ($M = 3$). In addition, there is a significant reduction in the computational complexity of the FW-PMWF method, which makes this algorithm feasible to be implemented in a hearing device.

It is important to remark that RECUP-MWF is not suitable for complexity reduction in WP-PMWF. In this case, linear solvers provide an efficient solution. In addition, although the RECUP-MWF provides significant complexity reduction in the FFT-based MWF and FW-based PMWF, PMWF-based implementations using WP and DWT and linear solvers are still the solutions with the smallest computational complexity.

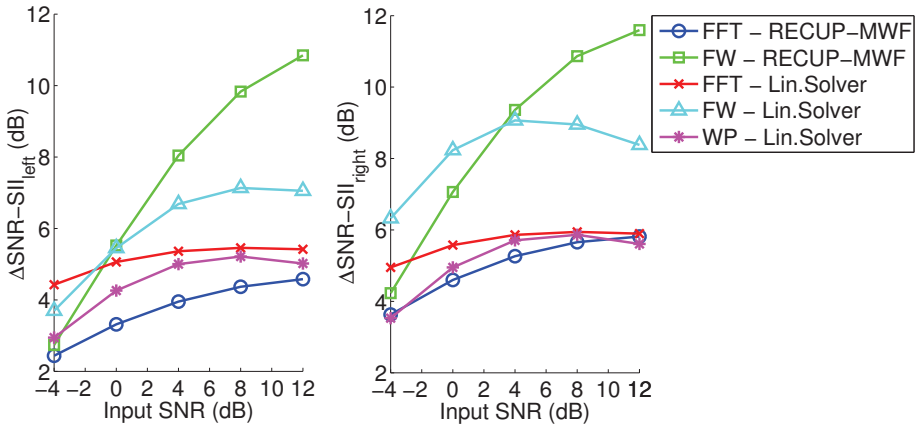

Figure 73: SNR improvement for different MWF implementations under babble noise scenario. MWF implementations include the RECUP-MWF implementation for FFT-MWF and FW-PMWF as well as FFT-MWF and WP-PMWF using linear solvers.

The performance of the RECUP-MWF implementations using FFT and FW are shown in the Figures 73 and 74. The RECUP-MWF implementation using FFT provides a performance nearly to the FFT-based implementation using linear solvers. On the other hand, the performance of RECUP-MWF for the FW-based implementation is better than the performance of the FW-based implementation using linear solvers, mainly at high input SNR. The latter is explained by the fact that the RECUP-MWF implementation introduces slight degradation in the speech quality. Hence, RECUP-MWF is an excellent alternative to implement the FW-PMWF method in a fixed-point DSP.

7.3 Reduction of Processing Artifacts in FFT-Based Processing

The implementation of most single-channel and multi-channel speech enhancement algorithms uses FFT-based block processing. In these algorithms, the frequency-domain filter weights are updated using a non-linear algorithm that invalidates the condition to avoid circular convolution, and then audible artifacts are present in the output. To minimize these processing artifacts, typical speech enhancement applications use a widely-known approach based on the following procedure: Overlapping by 50%, windowing, zero-padding, FFT, multiplication in the FFT-domain, IFFT, and overlap-add [47]. A graphical representation of this standard method, named standard windowed FFT convolution (SWFC) in this research, is shown in the Figure 75a. Another strategy employed to minimize artifacts is

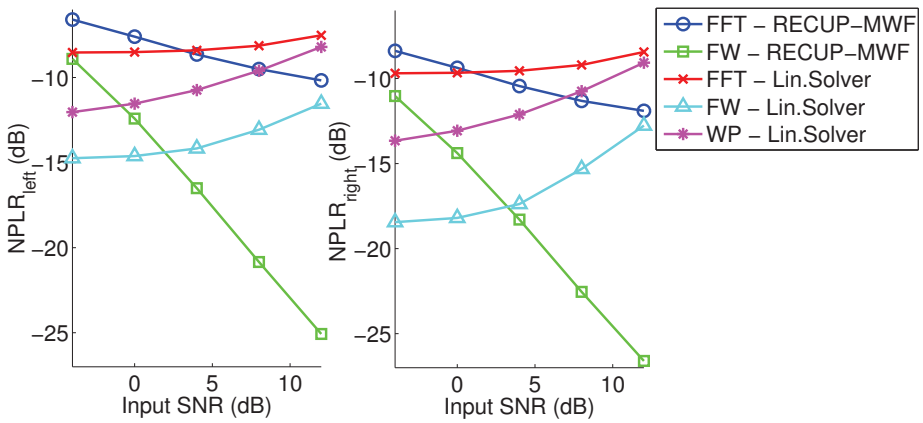

Figure 74: Noise reduction for different MWF implementations under babble noise scenario. MWF implementations include the RECUP-MWF implementation for FFT-MWF and FW-PMWF as well as FFT-MWF and WP-PMWF using linear solvers.

to use an analysis window to process the input blocks and a synthesis window to process the output blocks [8]. This approach, called double window FFT convolution (DWFC), is shown graphically in the Figure 76.

Even using the SWFC or DWFC approaches, audible artifacts may be present. To understand the source of this audible artifacts, we conducted a mathematical analysis of the SWFC and DWFC approaches, and proposed two artifact-free and distortion-free architectures that can be used for any speech enhancement algorithm based on FFT convolution [52, 53]. The following conclusions are discussed in [52, 53]:

- The output in SWFC differs from the expected output in some terms related exclusively to upper-half elements of the impulse response. This suggests that impulse responses to avoid artifacts in SWFC should be zero in the range $[N/2, N-1]$, where N is the FFT length.

- Artifacts in DWFC can be minimized, but not completely eliminated, using impulse responses whose elements in the range $[1/4N, 3/4N]$ are zero. This kind of impulse responses are common in speech enhancement algorithms and correspond to real, symmetric spectral gains. For these impulse responses, the DWFC approach offers lower distortion than SWFC, and for this reason DWFC is preferable for most implementations. In addition, a temporal analysis shows that DWFC assumes zero for

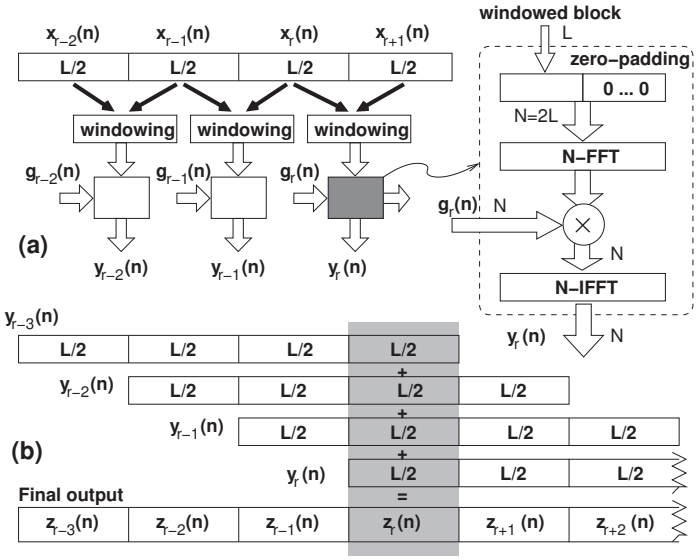

Figure 75: Block diagram for standard windowed FFT convolution (SWFC). (a) Processing of the input blocks, (b) Overlap-add of the output blocks. N is the FFT length, and $L = N/2$.

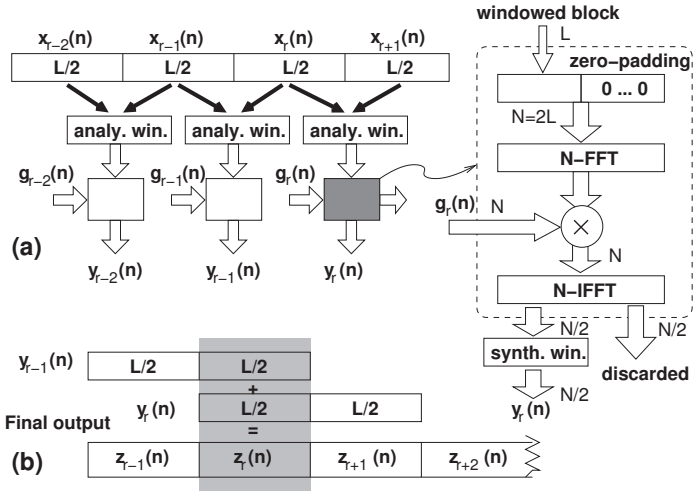

Figure 76: Block diagram for double window FFT convolution (DWFC). (a) Processing of the input blocks, (b) Overlap-add of the output blocks. N is the FFT length, and $L = N/2$.

the impulse response samples in the range $[1/4N, 3/4N]$. Hence, DWFC may introduce frequency-response distortions when the impulse response samples updated by the speech enhancement algorithm are not zero in this range.

- For SWFC, any tapered window[1] is unable to reduce artifacts, which suggests that windowing in SWFC is not enough to minimize artifacts.

- For DWFC, the synthesis, $w(n)$, and analysis, $v(n)$, windows must satisfy the condition[2]: $v(n)w(n) + w(n + L/2)v(n + L/2) = 1\ \forall n \in [0, L/2 - 1]$. But even using this condition, processing artifacts may be present in the output.

To remove processing artifacts and distortions on the frequency response, two approaches are proposed in this research. These artifact-free architectures are based on the extension of the frequency response (FEXT) and the splitting of the frequency response (FSPLT). The idea behind FEXT is the extension of the impulse response vector of length N to create a new frequency response vector of length $2N$ (Figure 77a). Thus, this algorithm can be seen as a SWFC algorithm using $2N$-FFTs rather than N-FFTs. On the other hand, FSPLT avoids the use of $2N$-FFTs by splitting the impulse response of length N into two impulse responses of length $N/2$. These impulses responses are zero-padded to get two new frequency responses of length N, and then two FFT convolutions are performed (Figure 77b).

A direct implementation of the block diagrams showed in the Figure 77 leads to expensive algorithms. A mathematical analysis of these structures allows to obtain more optimized block diagrams, which are presented in the Figures 78 and 79. We described a detailed mathematical derivation of these block diagrams in the references [52, 53]. In the Figures 78 and 79, F_N boxes describe FFT-like operations whose twiddle factor is denoted as a power of $F_N = \exp(-j2\pi/N)$. Therefore, F_N^{kn} describes an operator to compute an FFT of length N, F_N^{-kn} is an IFFT, $F_N^{(k+1/2)n}$ and $F_N^{(k-1/2)n}$ are modified FFT operations that can be computed in $\mathcal{O}(Nlog_2N)$ time.

[1] A tapered window $w(n)$ of length L satisfies the property $\sum_k w(n+kL) = const$. Triangular, Hamming and Hanning windows are examples of tapered windows.

[2] The mathematical proof is included in [53].

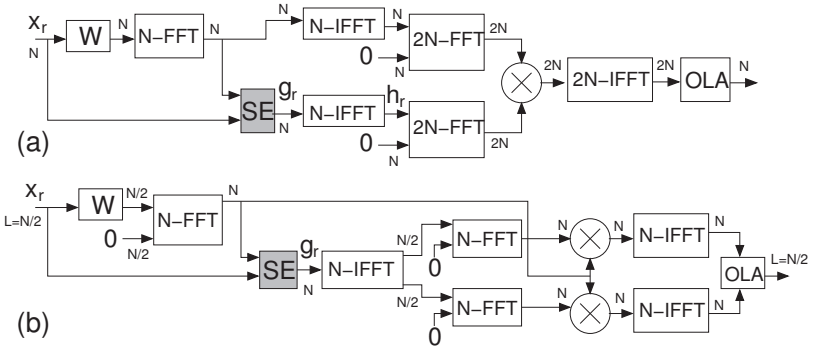

Figure 77: Principle of the two artifact-free FFT-convolution techniques applied to any speech enhancement algorithm (denoted by the shaded box SE). (a) FEXT and (b) FSPLT. W: Windowing, and OLA: Overlap-add.

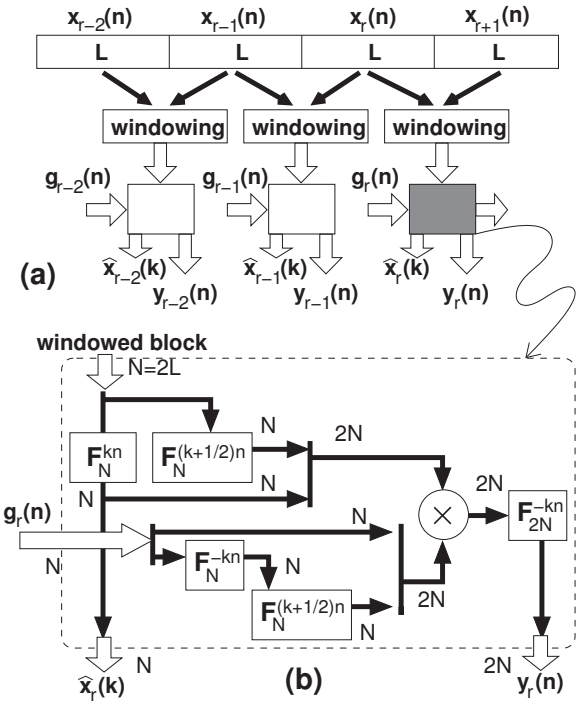

Figure 78: FFT convolution by frequency extension (FEXT).

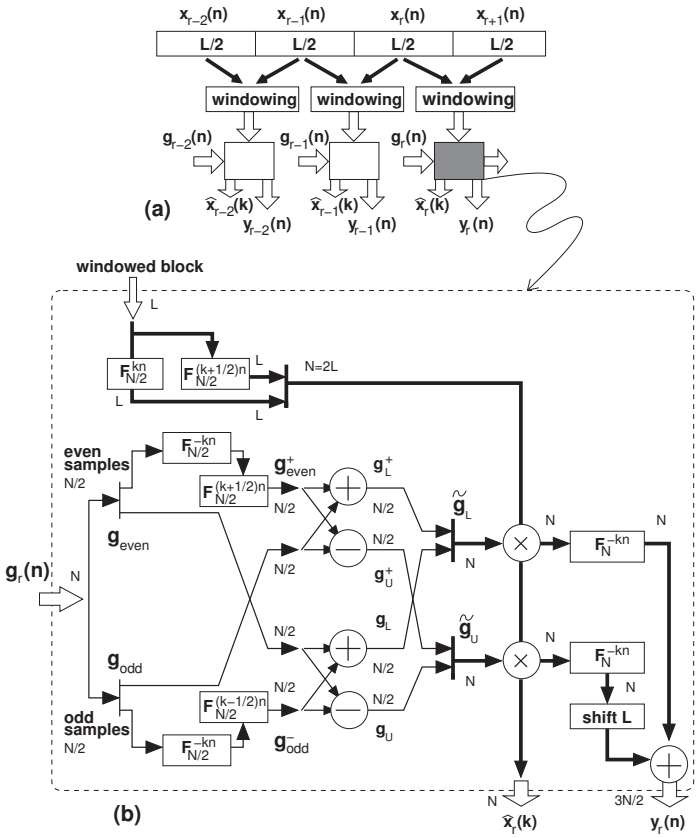

Figure 79: FFT convolution by frequency splitting (FSPLT).

Although both structures provide artifact-free block processing, there are two subtle differences between both approaches. First, to perform the processing with a weight vector of length N, FEXT uses input blocks of length N, whereas SWFC, DWFC and FSPLT use blocks of length $L = N/2$. However, the additional computational load required by FEXT for the frequency extension is compensated by processing twice the amount of data per block. Second, the computational cost in FSPLT is 2.5 times slower than SWFC, and FEXT is 1.6 times slower than SWFC. Hence, FEXT is preferable for most applications that require an artifact-free and distortion-free processing.

To verify the efficiency of the proposed methods in the removal of processing artifacts, two well-known speech enhancement algorithms that employ the FFT convolution are implemented using SWFC, DWFC, FEXT, and FSPLT, and a subjective test is conducted. These speech enhancement methods are the Wiener algorithm based on *a priori* SNR estimation (wiener_as) [85] and the minimum mean-square error log-spectral amplitude estimator algorithm (logMMSE) [17]. Subjects are asked to identify the presence of clicking sound and musical noise, and to rate the quality of each enhanced signal. The following is a summary of the results for the subjective test that we conducted in [53]:

- Audible artifacts identified as clicking sound are more noticeable in SWFC, and they can be removed using DWFC, FEXT and FSPLT.

- Audible artifacts identified as musical noise are strong in DWFC but they are not present in FEXT and FSPLT. Although DWFC can minimize the artifacts perceived as clicking sound, it introduces musical-noise artifacts. Musical noise is also present in SWFC but not as strong as in DWFC. This result is consistent with the mathematical framework, in which DWFC is shown to introduce more distortions than SWFC.

- Residual noise in FEXT and FSPLT preserves the structure and integrity of the original background noise. For example, for babble noise, residual noise is still distinguished as babble noise. But for DWFC and in lesser degree for SWFC, residual noise is distorted and heard as musical noise.

In summary, for any speech enhancement algorithm, subjects preferred the processing performed by FEXT and FSPLT. Taking into account the lower computation cost of FEXT compared to FSPLT, the absent of clicking sound and musical noise, and the good sound quality, FEXT is selected as the more suitable way to implement artifact-free FFT convolution.

Chapter VIII

CONCLUDING REMARKS

This research analyzed different binaural noise-reduction methods belonging to different categories: scene analysis, spectral subtraction, adaptive beamforming, multichannel Wiener filter (MWF), and blind source separation (BSS). From these existing methods, the MWF-based and BSS-based methods provide the best performance in terms of SNR improvement and sound quality. Since the implementation of these existing methods involve the usage of large number of operations per sample or block processing using large frame lengths, these algorithms are impractical for a digital hearing aid. *We found that by making design decisions based on an understanding of the human auditory system, it was possible to reduce latency, decrease the number of bands that were processed, decrease the required transmission rate, and improve noise reduction and speech quality and speech intelligibility.*

To reduce computational cost and latency, two methods were proposed in this research. These methods employ perceptual information to improve noise reduction, obtain better speech quality, and obtain feasible implementation strategies by removing unnecessary information from the perceptual viewpoint. In this case, the proposed implementations reduce computational cost, latency, and transmission bandwidth, and keep high noise reduction and speech quality. The first method, blind source separation and perceptual post-processing (BSS-PP), uses a BSS algorithm to get estimates of the speech and noise signals, and these estimates are used in a post processing stage to compute a set of time-domain gains that are used to cancel out the background noise. This post processing is based on a perceptual model that pushes down the noise level. Since the speech quality in BSS-PP is not good, a second method, perceptually-inspired MWF (PMWF), is proposed in this document. This method is based on MWF and replaces the FFT by a transformation that resembles the auditory fitlerbank. Two approaches were proposed for this transformation: wavelet packets (WP-PMWF) and frequency-warped filters (FW-PMWF). Different analysis showed that

WP-PMWF provides more benefits than the other methods, BSS-PP and FW-PMWF, because of the reduction in computational complexity, speech quality, and the feasibility to get an implementation that reduces the transmission bandwidth.

To reduce the transmission bandwidth, and so the power consumption, a method based on the WP-PMWF was proposed. The method uses a WP tree to decompose the input signal into sub-bands at different sampling rates. Then, binaural MWF is used only for the low-frequency sub-bands, while monaural MWF is used for the high-frequency sub-bands. This approach is based on our knowledge about the human perception of the localization cues, where interaural time difference (ITD) cues are more relevant for frequencies below 1.5kHz. Hence, to preserve the ITD cues, only the low-frequency sub-bands are required to be transmitted. As a result, the proposed method provides a performance close to a solution using full transmission of sub-bands and channels, a significant reduction in the transmission bandwidth, and an additional reduction in the computational complexity.

To reduce the computational complexity in WP-PMWF, the WP can be replaced by a discrete wavelet transform (DWT). Although the DWT does not match exactly an auditory filterbank, it provides a frequency analysis similar to the auditory filterbank, providing high resolution for the low-frequency sub-bands. This replacement provides similar performance to WP-PMWF and reduces the number of operation significantly, around 50%. The computational complexity in FW-PMWF can be reduced by using the recursive-update MWF algorithm (RECUP-MWF) proposed in this research. Although the computational cost of FW-PMWF using RECUP-MWF is reduced significantly compared to original version using linear solvers based on Cholesky decomposition, WP-PMWF is still the method with the lowest computational cost. Hence, among all different solutions proposed and analyzed in this study, the DWT-based PMWF implementation is the alternative with the lowest computational complexity, and its performance is nearly to the best method, WP-PMWF.

To estimate the second-order statistics, this research proposed a method based on a multichannel noise cross-PSD estimator and an adaptive trade-off parameter μ based on frame SNR. This method is a simplification of a MWF framework based on auditory masking thresholds and outperforms a VAD-based statistics estimation method, particularly to deal

with highly non-stationary environments.

Finally, to improve speech intelligibility, this research proposed a method based on MWF and binary masking (MWF-IDBM). This method provides significant improvement in terms of SNR, noise reduction, and speech intelligibility, and avoids the audible artifacts that are present in a standalone ideal binary masking (IDBM) method. The binary mask attempts to emulate the way how the brain is focused on one specific target signal. Although the MWF-IDBM solution is promising under ideal conditions, none of the mask estimation methods proposed in this research are sufficient to reach the speech intelligibility improvement achieved under ideal conditions.

As a final remark, we have learned that introducing modifications to the existing binaural noise-reduction methods based on our knowledge on perceptual properties is possible to achieve improvement in noise reduction, speech quality, and speech intelligibility, and simultaneously to satisfy the implementation constraints imposed by the hardware.

8.1 Contributions

This research resulted in the following contributions in a variety of publications:

- Identification of the BSS-based and MWF-based methods as the promising binaural noise-reduction methods to be used in a binaural hearing aid [51].
- Formulation of a mathematical framework to analyze the block-processing artifacts existing in single-channel and multi-channel speech enhancements methods implemented by FFT convolution. The standard overlap-add method is analyzed in [52], and a further extension to the double-window approach is analyzed in [53].
- Artifact-free and distortion-free architectures to perform FFT convolution [52, 53].
- Development of an on-line strategy to estimate the second-order statistics required by binaural MWF-based noise-reduction methods [56].
- A binaural noise-reduction method inspired by perceptual processing and BSS. This method is initially proposed in [62], and an extensive analysis and mathematical proofs about the preservation of localization cues are discussed in [63, 60].

- A MWF-based binaural noise-reduction method that uses perceptual processing rather than an FFT-based processing. A WP that resembles the auditory filterbank is proposed for this perceptual processing [54].

- A MWF-based noise-reduction method based on frequency-warped filters [59, 58].

- Development of a MWF-based binaural noise-reduction method to reduce efficiently the transmission bandwidth and computational complexity [55].

- Development of a method to reduce the number of operations involved in the implementation of the SDW-MWF framework and its application to FFT-based and FW-based MWF methods [58, 57].

- A proof of concept that a binaural noise-reduction method using a MWF framework and ideal binary masking (MWF-IDBM) is suitable to improve speech intelligibility and provide good noise reduction and sound quality in highly-noisy environments [61].

8.2 Suggestions for Future Research

- Although most implementation issues of the proposed binaural noise-reduction method based on BSS and perceptual post processing (BSS-PP) have been described in this research, a practical implementation requires to solve the following issues: a) To improve the sound quality, the dynamic range expansion performed by the post processing stage must include additional information to take into account a sound quality criteria, or use another perceptual model; b) To reduce the transmission bandwidth, it is necessary to develop distributive or reduced bandwidth BSS algorithms.

- Most processing in the BSS-PP method can be easily replaced by an analog processing except the BSS algorithm. A mixed-signal solution may reduce computational complexity and power consumption. To obtain a full-analog solution, analog BSS algorithms have to be developed.

- The WP-based MWF binaural noise-reduction method, WP-PMWF, has been identified as the most promising noise-reduction method for binaural hearing aids. Its

implementation in ultra low-power DSP architectures may require the exploration of hardware acceleration for the WP and weight computation.

- A test of WP-PMWF on real devices and its subjective validation by hearing-impaired people should be conducted to identify additional benefits of the proposed method.

- Although the proposed methods were not initially designed to deal with reverberant conditions, their performance under these environments is acceptable. Hence, their performance could be improvement by modifications in the mathematical framework to take into account the effect of reverberation.

- All methods proposed in this research were targeted for speech enhancement. Hence, some modifications are required for enhancement of other target signals such as music.

- A practical implementation of the method to improve speech intelligibility (MWF-IDBM) requires the development of reliable binary mask estimation algorithms.

- A simple speech intelligibility constraint was used to derive the MWF-IDBM framework. The robustness of MWF-IDBM against mask estimation errors could be mitigated by a mathematical framework derived from another speech intelligibility constraints, e.g., coherence speech intelligibility metrics.

- The MWF-IDBM framework showed to be useful under some scenarios at very low input SNR. This suggests that this processing can be enabled or disabled depending on the environmental condition. Hence, robust environmental classifiers have to be developed for an automatic enabling of this processing.

- Although the methods proposed in this research are targeted for binaural hearing aids, these ideas can be extended for other applications, e.g., noise reduction in mobile devices or automatic speech recognition systems.

- An analog model of the simplified recursive-update MWF method, RECUP-MWF, could be easily obtained. The latter opens the possibility to explore analog MWF implementations.

Appendix A

DERIVATION OF THE FREQUENCY-WARPED MWF FRAMEWORK (FW-PMWF)

Let $\tilde{y}_{m,k}$ the output of the all-pass filter at the m-th microphone and k-th tap given by

$$\tilde{y}_{m,k}(z) = A^k(z) y_m(z) \tag{45}$$

where $m = 1, ..., 2M$, with M the number of microphones per hearing aid, $k = 0, ..., K-1$, with K the number of taps in the frequency-warped FIR filter, and $A(z)$ is the transfer function of the all-pass filter given by

$$A(z) = \frac{z^{-1} - a}{1 - az^{-1}} \tag{46}$$

For a given time index n, the signals $\tilde{y}_{m,k}$ for all microphones and taps can be represented by a matrix of size $K \times 2M$,

$$\tilde{\boldsymbol{Y}}_n = \begin{bmatrix} \tilde{y}_{1,0}(n) & \tilde{y}_{2,0}(n) & & \tilde{y}_{2M,0}(n) \\ \tilde{y}_{1,1}(n) & \tilde{y}_{2,1}(n) & & \tilde{y}_{2M,1}(n) \\ \vdots & & \ddots & \cdots \\ \tilde{y}_{1,K-1}(n) & \tilde{y}_{2,K-1}(n) & \cdots & \tilde{y}_{2M,K-1}(n) \end{bmatrix} \tag{47}$$

Using this notation, the discrete Fourier transform, $\boldsymbol{F}$, of each column of $\tilde{\boldsymbol{Y}}_n$, $\hat{\boldsymbol{Y}}_n = \boldsymbol{F}\tilde{\boldsymbol{Y}}_n$, corresponds to the WDFT of the signals received at each microphone and time index n.

Assuming statistically independence between the target signal $x_m(n)$ and noise $v_m(n)$, i.e., $y_m(n) = x_m(n) + v_m(n)$, $\tilde{\boldsymbol{Y}}_n$ and $\hat{\boldsymbol{Y}}_n$ can be decomposed into signal and noise components as $\tilde{\boldsymbol{Y}}_n = \tilde{\boldsymbol{X}}_n + \tilde{\boldsymbol{V}}_n$ and $\hat{\boldsymbol{Y}}_n = \hat{\boldsymbol{X}}_n + \hat{\boldsymbol{V}}_n$, respectively.

In FW-PMWF, the outputs of the frequency-warped MWF for the left and right side are given by

$$z_L(n) \;\; = \;\; \sum_{m=1}^{2M} \sum_{k=0}^{K-1} w_L(m,k,n)\, \tilde{y}_{m,k}(n)$$

$$z_R(n) = \sum_{m=1}^{2M}\sum_{k=0}^{K-1} w_R(m,k,n)\tilde{y}_{m,k}(n)$$

where w_L and w_R are the filter weights. The above equations can be written in a more suitable way as

$$z_L(n) = \boldsymbol{w}_L^H(n)\, vec(\tilde{\boldsymbol{Y}}_n) \tag{48}$$

$$z_R(n) = \boldsymbol{w}_R^H(n)\, vec(\tilde{\boldsymbol{Y}}_n) \tag{49}$$

where $vec(.)$ represents the vectorization of the matrix $\tilde{\boldsymbol{Y}}_n$, and $\boldsymbol{w}_L(n)$ and $\boldsymbol{w}_R(n)$ are vectors of length $2MK$ given by

$$\begin{aligned}\boldsymbol{w}_L(n) &= [w_L^*(1,0,n)...w_L^*(1,K-1,n)...w_L^*(2M,0,n)...w_L^*(2M,K-1,n)]^H\\ \boldsymbol{w}_R(n) &= [w_R^*(1,0,n)...w_R^*(1,K-1,n)...w_R^*(2M,0,n)...w_R^*(2M,K-1,n)]^H\end{aligned}$$

Using the SDW-MWF framework, the filter weights are designed to minimize the MMSE between the speech components at the output of the frequency-warped FIR filters, $z_L^x(n)$ and $z_R^x(n)$, and the desired outputs, $x_L(n)$ and $x_R(n)$, subject to constraints in the output noise levels $z_L^v(n) < \theta$ and $z_R^v(n) < \theta$:

$$J_{SDW}(\boldsymbol{w}_L,\boldsymbol{w}_R) = \mathcal{E}\left\{\left\| \begin{array}{c} x_L(n) - z_L^x(n) \\ x_R(n) - z_R^x(n) \end{array}\right\|^2 + \mu \left\| \begin{array}{c} z_L^v(n) \\ z_R^v(n) \end{array}\right\|^2\right\} \tag{50}$$

When the time-warped information is used to compute the weights, i.e., using the matrix $\tilde{\boldsymbol{Y}}_n$, the coefficients at the left and right side are obtained after minimization of the cost function (50) with $x_L(n) = \boldsymbol{e}_L^H\tilde{\boldsymbol{X}}_n$, $x_R(n) = \boldsymbol{e}_R^H\tilde{\boldsymbol{X}}_n$, $z_L^x(n) = \boldsymbol{w}_L^H(n)\tilde{\boldsymbol{X}}_n$, $z_R^x(n) = \boldsymbol{w}_R^H(n)\tilde{\boldsymbol{X}}_n$, $z_L^v(n) = \boldsymbol{w}_L^H(n)\tilde{\boldsymbol{V}}_n$, and $z_R^v(n) = \boldsymbol{w}_R^H(n)\tilde{\boldsymbol{V}}_n$, i.e.,

$$J_{SDW}(\boldsymbol{w}_L,\boldsymbol{w}_R) = \mathcal{E}\left\{\left\| \begin{array}{c} (\boldsymbol{e}_L - \boldsymbol{w}_L)^H vec(\tilde{\boldsymbol{X}}_n) \\ (\boldsymbol{e}_R - \boldsymbol{w}_R)^H vec(\tilde{\boldsymbol{X}}_n) \end{array}\right\|^2 + \mu \left\| \begin{array}{c} \boldsymbol{w}_L^H vec(\tilde{\boldsymbol{V}}_n) \\ \boldsymbol{w}_R^H vec(\tilde{\boldsymbol{V}}_n) \end{array}\right\|^2\right\} \tag{51}$$

which yields to

$$\boldsymbol{w}_L(n) = \left(\boldsymbol{R}_{\tilde{X}}(n) + \mu\boldsymbol{R}_{\tilde{V}}(n)\right)^{-1}\boldsymbol{R}_{\tilde{X}}(n)\boldsymbol{e}_L \tag{52}$$

$$\boldsymbol{w}_R(n) = \left(\boldsymbol{R}_{\tilde{X}}(n) + \mu\boldsymbol{R}_{\tilde{V}}(n)\right)^{-1}\boldsymbol{R}_{\tilde{X}}(n)\boldsymbol{e}_R \tag{53}$$

where

$$\begin{aligned} \boldsymbol{R}_{\tilde{X}}(n) &\triangleq \mathcal{E}\left\{vec(\tilde{\boldsymbol{X}}_n)vec(\tilde{\boldsymbol{X}}_n)^H\right\} \\ \boldsymbol{R}_{\tilde{V}}(n) &\triangleq \mathcal{E}\left\{vec(\tilde{\boldsymbol{V}}_n)vec(\tilde{\boldsymbol{V}}_n)^H\right\} \end{aligned}$$

are the correlation matrices for the signal and noise components, respectively; and $\boldsymbol{e}_L$ and $\boldsymbol{e}_R$ are elementary vectors of length $2MK$ whose entry is one at the position of the reference microphone and tap $k = 0$.

On the other hand, when the frequency-warped information is used to compute the weights, i.e., using the matrix $\hat{\boldsymbol{Y}}_n$, the cost function is obtained by the following substitutions in (51): $vec(\tilde{\boldsymbol{X}}_n) = vec\left(\boldsymbol{F}^{-1}\hat{\boldsymbol{X}}_n\right) = \left(\boldsymbol{I}_{2M} \otimes \boldsymbol{F}^{-1}\right) vec(\hat{\boldsymbol{X}}_n)$, and $vec(\tilde{\boldsymbol{V}}_n) = \left(\boldsymbol{I}_{2M} \otimes \boldsymbol{F}^{-1}\right) vec(\hat{\boldsymbol{V}}_n)$, yielding to the following expressions to compute the coefficients

$$\boldsymbol{w}_L(n) = \left(\boldsymbol{I}_{2M} \otimes \boldsymbol{F}^{-1}\right) \left(\boldsymbol{R}_{\hat{X}}(n) + \mu \boldsymbol{R}_{\hat{V}}(n)\right)^{-1} \boldsymbol{R}_{\hat{X}}(n) \left(\boldsymbol{I}_{2M} \otimes \boldsymbol{F}\right) \boldsymbol{e}_L \quad (54)$$

$$\boldsymbol{w}_R(n) = \left(\boldsymbol{I}_{2M} \otimes \boldsymbol{F}^{-1}\right) \left(\boldsymbol{R}_{\hat{X}}(n) + \mu \boldsymbol{R}_{\hat{V}}(n)\right)^{-1} \boldsymbol{R}_{\hat{X}}(n) \left(\boldsymbol{I}_{2M} \otimes \boldsymbol{F}\right) \boldsymbol{e}_R \quad (55)$$

where

$$\begin{aligned} \boldsymbol{R}_{\hat{X}}(n) &\triangleq \mathcal{E}\left\{vec(\hat{\boldsymbol{X}}_n)vec(\hat{\boldsymbol{X}}_n)^H\right\} \\ \boldsymbol{R}_{\hat{V}}(n) &\triangleq \mathcal{E}\left\{vec(\hat{\boldsymbol{V}}_n)vec(\hat{\boldsymbol{V}}_n)^H\right\} \end{aligned}$$

and $\otimes$ denotes the tensor product.

Appendix B

DERIVATION OF THE MWF-IDBM FRAMEWORK

Weighs in MWF are derived from a minimization problem. In the SDW-MWF framework, the cost function is obtained from the minimization of the speech distortion at the reference microphone constrained to a given noise level[1]:

$$\min_{\boldsymbol{w}_L} \mathcal{E}\left\{ \left\| \begin{array}{c} (\boldsymbol{e}_L - \boldsymbol{w}_L)^H \boldsymbol{x} \\ (\boldsymbol{e}_R - \boldsymbol{w}_R)^H \boldsymbol{x} \end{array} \right\|^2 \right\} \; subject\, to \left\| \begin{array}{c} \boldsymbol{w}_L^H \boldsymbol{v} \\ \boldsymbol{w}_R^H \boldsymbol{v} \end{array} \right\|^2 < \Theta_{th}.$$

The idea of the MWF-IDBM method is to include an additional constraint related to the speech intelligibility. In [46], authors showed that speech intelligibility in the enhanced signal can be improved by including the constraint $|\hat{\boldsymbol{x}}|^2 \leq 4\,|\boldsymbol{x}|^2$, where $\hat{\boldsymbol{x}}$ is the enhanced signal. Hence, the proposed minimization problem is formulated as

$$\min_{\boldsymbol{w}_L} \mathcal{E}\left\{ \left\| \begin{array}{c} (\boldsymbol{e}_L - \boldsymbol{w}_L)^H \boldsymbol{x} \\ (\boldsymbol{e}_R - \boldsymbol{w}_R)^H \boldsymbol{x} \end{array} \right\|^2 \right\} \; subject\, to \left\| \begin{array}{c} \boldsymbol{w}_L^H \boldsymbol{v} \\ \boldsymbol{w}_R^H \boldsymbol{v} \end{array} \right\|^2 < \Theta_{th} \;\; and \;\; \mathcal{E}\,|\hat{\boldsymbol{x}}|^2 \leq 4\mathcal{E}\,|\boldsymbol{x}|^2 .$$

Thus, the filter coefficients are computed as

$$\boldsymbol{w}_L = \frac{1}{1+\delta}\left[\boldsymbol{R}_x + \frac{\mu+\delta}{1+\delta}\boldsymbol{R}_v\right]^{-1} \boldsymbol{R}_x \boldsymbol{e}_L \,,$$

where μ and δ are the Lagrangian operators related to the noise-level constraint and speech-intelligibility constraint, respectively. Since δ takes the values in the range $[0, \infty)$, the above equation can be reduced to

$$\boldsymbol{w}_L = g_L \left[\boldsymbol{R}_x + \mu' \boldsymbol{R}_v\right]^{-1} \boldsymbol{R}_x \boldsymbol{e}_L \,.$$

In the above equation, μ' plays a role similar to μ in SDW-MWF, i.e., to control the trade-off between speech distortion and noise reduction. On the other hand, g_L controls the speech intelligibility constraint. Since this parameter takes the values in the range $[0, 1]$, it can be seen as a binary mask.

[1]The derivation in this section if only for the left weights. The derivation for the right weights is identical.

Appendix C

DERIVATION OF RECURSIVE-UPDATE MWF (RECUP-MWF)

When the second-order statistics are updated using a VAD, the correlation matrices for the speech and noise are updated using the following rule that assumes stationary noise during the voiced segments:

if (noise-only segment)

$\boldsymbol{R}_v(f,l) = \lambda \boldsymbol{R}_v(f,l-1) + (1-\lambda)\boldsymbol{y}(f,l)\boldsymbol{y}^H(f,l)$

$\boldsymbol{R}_x(f,l) = \lambda \boldsymbol{R}_x(f,l-1)$

else //voiced-segment

$\boldsymbol{R}_y(f,l) = \lambda \boldsymbol{R}_y(f,l-1) + (1-\lambda)\boldsymbol{y}(f,l)\boldsymbol{y}^H(f,l)$

$\boldsymbol{R}_x(f,l) = \boldsymbol{R}_y(f,l-1) - \boldsymbol{R}_v(f,l-1)$

$\boldsymbol{R}_v(f,l) = \boldsymbol{R}_v(f,l-1)$

end

In the above equation, λ is a forgetting factor whose value is close to 1, and $\boldsymbol{y}$ is the input vector of length $2M \times 1$. Therefore, the correlation matrices $\boldsymbol{R}_x$ and $\boldsymbol{R}_v$ are updated recursively based on the addition of rank-one matrices $\boldsymbol{y}\boldsymbol{y}^H$. In this sense, an efficient algorithm to compute the MWF weights can be obtained by replacing $\boldsymbol{R}_x$ and $\boldsymbol{R}_v$ given by the above rules into the equations (56) and (57):

$$\boldsymbol{w}_L(f,l) = \boldsymbol{R}^{-1}(f,l)\boldsymbol{R}_x(f,l)\boldsymbol{e}_L \quad (56)$$

$$\boldsymbol{w}_R(f,l) = \boldsymbol{R}^{-1}(f,l)\boldsymbol{R}_x(f,l)\boldsymbol{e}_R \quad (57)$$

$$\boldsymbol{R}(f,l) = \boldsymbol{R}_x(f,l) + \mu \boldsymbol{R}_v(f,l) \quad (58)$$

Thus, it is possible obtain a recursive-update algorithm to compute the inverse of the correlation matrix $\boldsymbol{R}(f,l)$. For a voiced segment, the matrix $\boldsymbol{R}(f,l)^{-1}$ has the form:

$$\boldsymbol{R}(f,l)^{-1} = [\boldsymbol{R}_y(f,l) - \boldsymbol{R}_v(f,l-1) + \mu \boldsymbol{R}_v(f,l)]^{-1}$$

$$= \left[\lambda \boldsymbol{R}_x(f,l-1) + (\mu - 1 + \lambda)\boldsymbol{R}_v(f,l-1) + (1-\lambda)\boldsymbol{y}(f,l)\boldsymbol{y}^H(f,l)\right]^{-1} \quad (59)$$

Since $\lambda \approx 1$, the term $\lambda \boldsymbol{R}_x(f,l-1) + (\mu - 1 + \lambda)\boldsymbol{R}_v(f,l-1)$ can be approximated as $\alpha(\boldsymbol{R}_x(f,l-1) + \mu \boldsymbol{R}_v(f,l-1)) = \alpha \boldsymbol{R}(f,l-1)$, where α is a forgetting factor whose value is close to 1. This value must be close to 1 to ensure the stability of the algorithm. A value of $\alpha = 0.995$ is chosen for all experiments. Thus, using the matrix inversion lemma,

$$\begin{aligned} \boldsymbol{R}(f,l)^{-1} &= \left[\alpha \boldsymbol{R}(f,l-1) + (1-\lambda)\boldsymbol{y}(f,l)\boldsymbol{y}^H(f,l)\right]^{-1} \\ &= \frac{1}{\alpha}\left[\boldsymbol{R}^{-1}(f,l-1) - (1-\lambda)\beta(f,l)\boldsymbol{q}(f,l)\boldsymbol{q}^H(f,l)\right] \end{aligned} \quad (60)$$

where

$$\boldsymbol{q}(f,l) = \boldsymbol{R}^{-1}(f,l-1)\boldsymbol{y}(f,l) \quad (61)$$

and

$$\beta(f,l) = \frac{1}{\alpha + (1-\lambda)\boldsymbol{y}^H(f,l)\boldsymbol{q}(f,l)}. \quad (62)$$

For a noise-only segment, the matrix $\boldsymbol{R}(f,l)^{-1}$ differs from (59):

$$\begin{aligned} \boldsymbol{R}(f,l)^{-1} &= \left[\lambda \boldsymbol{R}_x(f,l-1) + \mu \boldsymbol{R}_v(f,l)\right]^{-1} \\ &= \left[\lambda \boldsymbol{R}_x(f,l-1) + \mu\lambda \boldsymbol{R}_v(f,l-1) + \mu(1-\lambda)\boldsymbol{y}(f,l)\boldsymbol{y}^H(f,l)\right]^{-1} \end{aligned}$$

Again, since $\lambda \approx 1$, the term $\lambda \boldsymbol{R}_x(f,l-1) + \mu\lambda \boldsymbol{R}_v(f,l-1)$ can be approximated as $\alpha(\boldsymbol{R}_x(f,l-1) + \mu \boldsymbol{R}_v(f,l-1)) = \alpha \boldsymbol{R}(f,l-1)$, and after applying the matrix inversion lemma,

$$\begin{aligned} \boldsymbol{R}(f,l)^{-1} &= \left[\alpha \boldsymbol{R}(f,l-1) + \mu(1-\lambda)\boldsymbol{y}(f,l)\boldsymbol{y}^H(f,l)\right]^{-1} \\ &= \frac{1}{\alpha}\left[\boldsymbol{R}^{-1}(f,l-1) - \mu(1-\lambda)\beta'(f,l)\boldsymbol{q}(f,l)\boldsymbol{q}^H(f,l)\right] \end{aligned} \quad (63)$$

where $\boldsymbol{q}(f,l)$ is defined as in (61) and

$$\beta'(f,l) = \frac{1}{\alpha + \mu(1-\lambda)\boldsymbol{y}^H(f,l)\boldsymbol{q}(f,l)} \quad (64)$$

Note that the only difference for the definitions of the inverse matrix $\boldsymbol{R}^{-1}$ and β for voiced and unvoiced segments is in the use of the constant μ.

The number of operations involved in the weight and output computation can be reduced by simple substitutions:

//Initialization

$\boldsymbol{R}^{-1}(f,0) = \boldsymbol{I}$

//Processing

$\boldsymbol{q}(f,l) = \boldsymbol{R}^{-1}(f,l-1)\boldsymbol{y}(f,l)$

if (voiced segment)

$\boldsymbol{r}_{x_L}(f,l) = \lambda \boldsymbol{r}_{x_L}(f,l-1) + (1-\lambda) y_L^*(f,l)\boldsymbol{y}(f,l)$

$\boldsymbol{r}_{x_R}(f,l) = \lambda \boldsymbol{r}_{x_R}(f,l-1) + (1-\lambda) y_R^*(f,l)\boldsymbol{y}(f,l)$

$\mu' = 1-\lambda$

else //unvoiced segment

$\boldsymbol{r}_{x_L}(f,l) = \lambda \boldsymbol{r}_{x_L}(f,l-1)$

$\boldsymbol{r}_{x_R}(f,l) = \lambda \boldsymbol{r}_{x_R}(f,l-1)$

$\mu' = \mu(1-\lambda)$

end

$\beta(f,l) = \frac{1}{\alpha + \mu' \boldsymbol{y}^H(f,l)\boldsymbol{q}(f,l)}$

$\boldsymbol{R}(f,l)^{-1} = \frac{1}{\alpha}\left[\boldsymbol{R}^{-1}(f,l-1) - \mu'\beta(f,l)\boldsymbol{q}(f,l)\boldsymbol{q}^H(f,l)\right]$

$z_L(f,l) = \beta(f,l)\boldsymbol{r}_{x_L}^H(f,l)\boldsymbol{q}(f,l)$

$z_R(f,l) = \beta(f,l)\boldsymbol{r}_{x_R}^H(f,l)\boldsymbol{q}(f,l)$

Figure 80: Recursive-Update MWF.

$$z_L(f,l) = \boldsymbol{w}_L^H(f,l)\boldsymbol{y}(f,l) = \boldsymbol{r}_{x_L}^H(f,l)\boldsymbol{R}^{-1}(f,l)\boldsymbol{y}(f,l)$$

where $\boldsymbol{r}_{x_L}(f,l) = \boldsymbol{R}_x(f,l)\boldsymbol{e}_L$. For a voiced segment, replacing $\boldsymbol{R}^{-1}(f,l)$ by (60), the output becomes,

$$z_L(f,l) = \beta(f,l)\boldsymbol{r}_{x_L}^H(f,l)\boldsymbol{q}(f,l) \tag{65}$$

a similar equation is obtained for the unvoiced segments but using β' instead of β.

The Figure 80 includes a summary of the proposed algorithm. Since this algorithm uses a recursive update for inverse correlation matrix, the algorithm is called recursive-update MWF or RECUP-MWF.

Bibliography

[1] Aichner, R., Buchner, H., Zourub, M., and Kellermann, W., "Multi-channel source separation preserving spatial information," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, (Honolulu, USA), pp. I–5–I–8, 2007.

[2] Brungart, D. S., Chang, P. S., Simpson, B. D., and Wang, D., "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation," *The Journal of the Acoustical Society of America*, vol. 120, no. 6, pp. 4007–4018, 2006.

[3] Chisaki, Y., Matsuo, K., Hagiwara, K., Nakashima, H., and Usagawa, T., "Real-time processing using the frequency domain binaural model," *Applied Acoustics*, vol. 68, no. 8, pp. 923 – 938, 2007.

[4] Cooke, M., Green, P., Josifovski, L., and Vizinho, A., "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Communication*, vol. 34, no. 3, pp. 267–285, 2001.

[5] Cornelis, B., Doclo, S., Van dan Bogaert, T., Moonen, M., and Wouters, J., "Theoretical analysis of binaural multimicrophone noise reduction techniques," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, pp. 342 – 355, Feb. 2010.

[6] Cornelis, B., Moonen, M., and Wouters, J., "Performance analysis of multichannel Wiener filter-based noise reduction in hearing aids under second order statistics estimation errors," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1368 –1381, 2011.

[7] Cornells, B., Moonen, M., and Wouters, J., "A VAD-robust multichannel Wiener filter algorithm for noise reduction in hearing aids," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (Prague, Czech Republic), pp. 281 –284, May 2011.

[8] Crochiere, R., "A weighted overlap-add method of short-time fourier analysis/synthesis," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, pp. 99–102, Feb. 1980.

[9] Department of Electrical and Electronic Engineering, Imperial College, London, "VOICEBOX: Speech processing toolbox for MATLAB," 2003. http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html.

[10] Doclo, S., Dong, R., Klasen, T., Wouters, J., Haykin, S., and Moonen, M., "Extension of the multi-channel Wiener filter with ITD cues for noise reduction in binaural hearing aids," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, (New Paltz, USA), pp. 70–73, Oct. 2005.

[11] Doclo, S. and Moonen, M., "GSVD-based optimal filtering for single and multimicrophone speech enhancement," *IEEE Transactions on Signal Processing*, vol. 50, pp. 2230 – 2244, Sept. 2002.

[12] Doclo, S. and Moonen, M., "Multimicrophone noise reduction using recursive GSVD-based optimal filtering with ANC postprocessing stage," *IEEE Transactions on Speech and Audio Processing*, vol. 13, pp. 53 – 69, Jan. 2005.

[13] Doclo, S., Moonen, M., Van den Bogaert, T., and Wouters, J., "Reduced-bandwidth and distributed MWF-based noise reduction algorithms for binaural hearing aids," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, pp. 38–51, Jan. 2009.

[14] Doclo, S., Spriet, A., Wouters, J., and Moonen, M., "Speech distortion weighted multichannel Wiener filtering techniques for noise reduction," in *Speech Enhancement* (Benesty, J., Chen, J., and Makino, S., eds.), pp. 199 – 228, Berlin: Springer, 2005.

[15] Doclo, S., Van den Bogaert, T., Wouters, J., and Moonen, M., "Comparison of reduced-bandwidth MWF-based noise reduction algorithms for binaural hearing aids," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, (New Paltz, USA), pp. 223–226, Oct. 2007.

[16] Ephraim, Y. and Malah, D., "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, pp. 1109–1121, Dec. 1984.

[17] Ephraim, Y. and Malah, D., "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 23, no. 2, pp. 443–445, 1985.

[18] European Telecommunications Standards Institute (ETSI), "Technical specification ETSI TS 126 077 V10.0.0 (2011-04)," 2011. http://www.etsi.org/deliver/etsi-ts/126000-126099/126077/10.00.00-60/.

[19] Gardner, B. and Martin, K., "HRTF measurements of a KEMAR dummy-head microphone," Tech. Rep. 280, MIT Media Lab Perceptual Computing, 1994. http://sound.media.mit.edu/KEMAR.html.

[20] Goetze, S., Rohdenburg, T., Hohmann, V., Kollmeier, B., and Kammeyer, K., "Direction of arrival estimation based on the dual delay line approach for binaural hearing aid microphone arrays," in *Proc. International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*, (Xiamen, China), pp. 84–87, Dec. 2007.

[21] Greenberg, J. E., Peterson, P. M., and Zurek, P. M., "Intelligibility-weighted measures of speech-to-interference ratio and speech system performance," *The Journal of the Acoustical Society of America*, vol. 94, no. 5, pp. 3009–3010, 1993.

[22] Hu, G. and Wang, D., "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Transactions on Neural Network*, vol. 15, pp. 1135 – 1150, sept. 2004.

[23] Huber, R. and Kollmeier, B., "PEMO-Q a new method for objective audio quality assessment using a model of auditory perception," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, pp. 1902–1911, Nov. 2006.

[24] Humes, L. E., Christensen, L., Thomas, T., Bess, F. H., Hedley-Williams, A., and Bentler, R., "A comparison of the aided performance and benefit provided by a linear and a two-channel wide dynamic range compression hearing aid," *Journal of Speech Language and Hearing Research*, vol. 42, no. 1, pp. 65–79, 1999.

[25] IEEE Subcommittee, "IEEE recommended practice for speech quality measurements," *IEEE Transactions on Audio Electroacoustic*, pp. 225–246, 1969.

[26] ITU-R, "Recommendation BS.1534-1: Method for the subjective assessment of intermediate quality levels of coding systems," 2003.

[27] ITU-T, "Recommendation P.862: Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," 2001.

[28] ITU-T, "Recommendation P.835: Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm," 2003.

[29] Jeub, M., Schafer, M., and Vary, P., "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *Proc. International Conference on Digital Signal Processing*, pp. 1 –5, 2009.

[30] Johnston, J., "Estimation of perceptual entropy using noise masking criteria," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP)*, pp. 2524 –2527 vol.5, Apr 1988.

[31] Kamkar-Parsi, A. and Bouchard, M., "Improved noise power spectrum density estimation for binaural hearing aids operating in a diffuse noise field environment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, pp. 521–533, May 2009.

[32] Karmakar, A., Kumar, A., and Patney, R., "Design of optimal wavelet packet trees based on auditory perception criterion," *IEEE Signal Processing Letters*, vol. 14, no. 4, pp. 240 –243, 2007.

[33] Kates, J. M., *Digital Hearing Aids*. San Diego, CA: Plural Publishing, 2008.

[34] Kates, J. M. and Arehart, K. H., "Coherence and the speech intelligibility index," *The Journal of the Acoustical Society of America*, vol. 117, no. 4, pp. 2224–2237, 2005.

[35] Kates, J. M. and Arehart, K. H., "Multichannel dynamic-range compression using digital frequency warping," *EURASIP Journal on Applied Signal Processing*, vol. 2005, no. 18, pp. 3003–3014, 2005.

[36] Kim, G., "Multi-microphone interference suppression using the principal subspace modification and its application to speech recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5508 –5511, May 2011.

[37] Kim, G. and Cho, N. I., "QRD-based optimal filtering for multichannel speech enhancement using time varying smoothing factor," in *Proc. International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*, pp. 37 – 40, Dec. 2005.

[38] Kim, G., Lu, Y., Hu, Y., and Loizou, P. C., "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," *The Journal of the Acoustical Society of America*, vol. 126, no. 3, pp. 1486–1494, 2009.

[39] Klasen, T., Moonen, M., Van den Bogaert, T., and Wouters, J., "Preservation of interaural time delay for binaural hearing aids through multi-channel Wiener filtering based noise reduction," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, Proceedings (ICASSP)*, vol. 3, (Philadelphia, USA), pp. iii/29–iii/32, Mar. 2005.

[40] Klasen, T., Van den Bogaert, T., Moonen, M., and Wouters, J., "Binaural noise reduction algorithms for hearing aids that preserve interaural time delay cues," *IEEE Transactions on Signal Processing*, vol. 55, pp. 1579–1585, Apr. 2007.

[41] Kocinski, J., "Speech intelligibility improvement using convolutive blind source separation assisted by denoising algorithms," *Speech Communication*, vol. 50, no. 1, pp. 29 – 37, 2008.

[42] Li, J., Akagi, M., and Suzuki, Y., "Extension of the two-microphone noise reduction method for binaural hearing aids," in *Proc. International Conference on Audio, Language and Image Processing (ICALIP)*, (Shanghai, China), pp. 97–101, 2008.

[43] Li, J., Sakamoto, S., Hongo, S., Akagi, M., and Suzuki, Y., "Two-stage binaural speech enhancement with Wiener filter based on equalization-cancellation model," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, (New Paltz, USA), pp. 133 –136, Oct. 2009.

[44] Li, J., Sakamoto, S., Hongo, S., Akagi, M., and Suzuki, Y., "Two-stage binaural speech enhancement with Wiener filter for high-quality speech communication," *Speech Communication*, vol. 53, no. 5, pp. 677 – 689, 2011.

[45] Li, N. and Loizou, P. C., "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," *The Journal of the Acoustical Society of America*, vol. 123, no. 3, pp. 1673–1682, 2008.

[46] Loizou, P. and Kim, G., "Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, pp. 47 –56, Jan. 2011.

[47] Loizou, P. C., *Speech Enhancement. Theory and Practice*. Boca Raton, FL: CRC Press, 2007.

[48] Lotter, T. and Vary, P., "Dual-channel speech enhancement by superdirective beamforming," *EURASIP Journal on Applied Signal Processing*, vol. 2006, pp. 175–175, 2006.

[49] Luo, F.-L., Yang, J., Pavlovic, C., and Nehorai, A., "Adaptive null-forming scheme in digital hearing aids," *IEEE Transactions on Signal Processing*, vol. 50, pp. 1583–1590, July 2002.

[50] Makur, A. and Mitra, S., "Warped discrete-fourier transform: Theory and applications," *Circuits and Systems I: Fundamental Theory and Applications, IEEE Transactions on*, vol. 48, pp. 1086 –1093, Sep. 2001.

[51] Marin-Hurtado, J. I. and Anderson, D. V., "Comparative study of eleven noise reduction techniques for binaural hearing aids," in *Proc. International Hearing Aid Research Conference (IHCON)*, (Lake Tahoe, USA), p. 65, Aug. 2010.

[52] Marin-Hurtado, J. I. and Anderson, D. V., "Distortions in speech enhancement due to block processing," in *Proc. IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, (Dallas, USA), pp. 4774 –4777, 2010.

[53] Marin-Hurtado, J. I. and Anderson, D. V., "FFT-based block processing in speech enhancement: Potential artifacts and solutions," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, pp. 2527 –2537, Nov. 2011.

[54] Marin-Hurtado, J. I. and Anderson, D. V., "Perceptually-inspired processing for multichannel Wiener filter," in *Proc. Interspeech 2011*, vol. 1, (Florence, Italy), pp. 1777–1780, Aug. 2011.

[55] Marin-Hurtado, J. I. and Anderson, D. V., "Reduced-bandwidth and low-complexity multichannel Wiener filter for binaural hearing aids," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, (New Paltz, NY), pp. 85 –88, Oct. 2011.

[56] Marin-Hurtado, J. I. and Anderson, D. V., "Robust non-VAD implementation of multichannel Wiener filter for binaural noise reduction," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, (New Paltz, NY), pp. 341 –344, Oct. 2011.

[57] Marin-Hurtado, J. I. and Anderson, D. V., "An analog model to implement a multichannel Wiener filter," *To be published in IEEE Transactions on Circuits and Systems II*, 2012.

[58] Marin-Hurtado, J. I. and Anderson, D. V., "Binaural noise reduction using frequency-warped filters and multichannel Wiener filter," *To be published in IEEE Transactions on Audio, Speech and Language Processing*, 2012.

[59] Marin-Hurtado, J. I. and Anderson, D. V., "Binaural noise reduction using frequency-warped FIR filters," in *Submitted to Interspeech 2012*, (Portland, OR), Sep. 2012.

[60] Marin-Hurtado, J. I. and Anderson, D. V., "Preservation of localization cues in BSS-based noise reduction: Application in binaural hearing aids," in *Independent Component Analysis for Audio and Biosignal applications* (R. Naik, G., ed.), Intech, June 2012.

[61] Marin-Hurtado, J. I. and Anderson, D. V., "Speech intelligibility improvement in multichannel wiener filters by binary masking," *To be published in IEEE Transactions on Audio, Speech and Language Processing*, 2012.

[62] Marin-Hurtado, J. I., Parikh, D. N., and Anderson, D. V., "Binaural noise-reduction method based on blind source separation and perceptual post processing," in *Proc. Interspeech 2011*, vol. 1, (Florence, Italy), pp. 217–220, Aug. 2011.

[63] Marin-Hurtado, J. I., Parkih, D. N., and Anderson, D. V., "Perceptually inspired noise-reduction method for binaural hearing aids," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, pp. 1372–1382, May 2012.

[64] Moore, B. C. J., "Binaural sharing of audio signals: Prospective benefits and limitations," *The Hearing Journal*, vol. 60, no. 11, pp. 46–48, 2007.

[65] Moore, B. C., *An introduction to the Psychology of Hearing*. London: Elsevier, 2007.

[66] Neuman, A. C., Bakke, M. H., Hellman, S., and Levitt, H., "Effect of compression ratio in a slow-acting compression hearing aid: Paired-comparison judgments of quality," *The Journal of the Acoustical Society of America*, vol. 96, no. 3, pp. 1471–1478, 1994.

[67] Ngo, K., Spriet, A., Moonen, M., Wouters, J., and Jensen, S. H., "Variable speech distortion weighted multichannel Wiener filter based on soft output voice activity detection for noise reduction in hearing aids," in *Proc. 11th International Workshop on Acoustic Echo and Noise Control (IWAENC)*, (Seattle, USA), Sept. 2008.

[68] Nishimura, R., Suzuki, Y., Tsukui, S., and Asano, F., "Array signal processing with two outputs preserving binaural information," *Applied Acoustics*, vol. 65, no. 7, pp. 657 – 672, 2004.

[69] Noohi, T. and Kahaei, M., "Residual cross-talk suppression for convolutive blind source separation," in *Proc. International Conference on Computer Engineering Technology (ICCET)*, vol. 1, pp. V1–543 –V1–547, 2010.

[70] Oticon, "Oticon EPOQ," 2009. http://www.oticonusa.com/Oticon/ Professionals/professional_products/Epoq/epoq_xw_vs_w.html.

[71] Parfieniuk, M. and Petrovsky, A., "Warped DFT as the basis for psychoacoustic model," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 4, pp. iv–185 – iv–188 vol.4, May 2004.

[72] Parikh, D. and Anderson, D., "Blind source separation with perceptual post processing," in *Proc. IEEE Workshop on Digital Signal Processing and Signal Processing Education (DSP/SPE)*, pp. 321 –325, Jan. 2011.

[73] Phonak, "Phonak zoomcontrol," 2009. http://www.exelia.phonak.com/en/ features-and-functions/control/zoomcontrol/.

[74] Rahmani, M., Akbari, A., and Ayad, B., "An iterative noise cross-PSD estimation for two-microphone speech enhancement," *Applied Acoustics*, vol. 70, no. 3, pp. 514 – 521, 2009.

[75] Reindl, K., Zheng, Y., and Kellermann, W., "Speech enhancement for binaural hearing aids based on blind source separation," in *Proc. International Symposium on Commununications, Control and Signal Processing (ISCCSP)*, pp. 1 –6, March 2010.

[76] Richards, V., Moore, B., and Launer, S., "Potential benefits of across-aid communication for bilaterally aided people: listening in a car," *International Journal of Audiology*, vol. 45, no. 3, pp. 182 – 189, 2006.

[77] Ris, C. and Dupont, S., "Assessing local noise level estimation methods: Application to noise robust ASR," *Speech Communication*, vol. 34, no. 1-2, pp. 141 – 158, 2001.

[78] Rohdenburg, T., Goetze, S., Hohmann, V., Kammeyer, K., and Kollmeier, B., "Objective perceptual quality assessment for self-steering binaural hearing aid microphone arrays," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (Las Vegas, USA), pp. 2449–2452, 2008.

[79] Rohdenburg, T., Hohmann, V., and Kollmeier, B., "Robustness analysis of binaural hearing aid beamformer algorithms by means of objective perceptual quality

measures," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, (New Paltz, USA), pp. 315–318, Oct. 2007.

[80] Roman, N., Srinivasan, S., and Wang, D., "Binaural segregation in multisource reverberant environments," *The Journal of the Acoustical Society of America*, vol. 120, no. 6, pp. 4040–4051, 2006.

[81] Rombouts, G. and Moonen, M., "Fast QRD-lattice-based unconstrained optimal filtering for acoustic noise reduction," *IEEE Transactions on Speech and Audio Processing*, vol. 13, pp. 1130 – 1143, Nov. 2005.

[82] Rombouts, G. and Moonen, M., "QRD-based unconstrained optimal filtering for acoustic noise reduction," *Signal Processing*, vol. 83, no. 9, pp. 1889 – 1904, 2003.

[83] Roy, O. and Vetterli, M., "Rate-constrained collaborative noise reduction for wireless hearing aids," *IEEE Transactions on Signal Processing*, vol. 57, pp. 645–657, Feb. 2009.

[84] RWTH Aachen University, "Aachen impulse response (AIR) database - version 1.2," 2010. http://www.ind.rwth-aachen.de/AIR.

[85] Scalart, P. and Filho, J., "Speech enhancement based on a priori signal to noise estimation," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, Proceedings (ICASSP)*, vol. 2, (Atlanta, USA), pp. 629–632, May 1996.

[86] Seelamantula, C. S. and Sreenivas, T. V., "Blocking artifacts in speech/audio: Dynamic auditory model-based characterization and optimal time-frequency smoothing," *Signal Processing*, vol. 89, no. 4, pp. 523 – 531, 2009.

[87] Siemens, "e2e wireless: Technical overview," 2004. http://mysiemens. siemens-hearing.com/admin/documents/10053817e2eWirelessOverview.pdf.

[88] Smith, J.O., I. and Abel, J., "Bark and ERB bilinear transforms," *IEEE Transactions on Speech and Audio Processing*, vol. 7, pp. 697 –708, nov 1999.

[89] Smith, P., Davis, A., Day, J., Unwin, S., Day, G., and Chalupper, J., "Real-world preferences for linked bilateral processing," *The Hearing Journal*, vol. 61, no. 7, pp. 33–38, 2008.

[90] Sockalingam, R., Holmberg, M., Eneroth, K., and Shulte, M., "Binaural hearing aid communication shown to improve sound quality and localization," *The Hearing Journal*, vol. 62, no. 10, pp. 46–47, 2009.

[91] Souden, M., Benesty, J., and Affes, S., "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, pp. 260 –276, Feb. 2010.

[92] Spriet, A., Moonen, M., and Wouters, J., "Stochastic gradient implementation of spatially preprocessed multi-channel Wiener filtering for noise reduction in hearing aids," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 4, pp. iv–57 – iv–60 vol.4, May 2004.

[93] Spriet, A., Moonen, M., and Wouters, J., "Stochastic gradient-based implementation of spatially preprocessed speech distortion weighted multichannel Wiener filtering for noise reduction in hearing aids," *IEEE Transactions on Signal Processing*, vol. 53, pp. 911 – 925, March 2005.

[94] Srinivasan, S., "Low-bandwidth binaural beamforming," *Electronics Letters*, vol. 44, no. 22, pp. 1292–1293, 2008.

[95] Srinivasan, S., "Noise reduction in binaural hearing aids: Analyzing the benefit over monaural systems," *The Journal of the Acoustical Society of America*, vol. 124, no. 6, pp. EL353–EL359, 2008.

[96] Takahashi, Y., Takatani, T., Osako, K., Saruwatari, H., and Shikano, K., "Blind spatial subtraction array for speech enhancement in noisy environment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, pp. 650 –664, May 2009.

[97] Van den Bogaert, T., Doclo, S., Wouters, J., and Moonen, M., "The effect of multimicrophone noise reduction systems on sound source localization by users of binaural hearing aids," *The Journal of the Acoustical Society of America*, vol. 124, no. 1, pp. 484–497, 2008.

[98] Van den Bogaert, T., Doclo, S., Wouters, J., and Moonen, M., "Speech enhancement with multichannel Wiener filter techniques in multimicrophone binaural hearing aids," *The Journal of the Acoustical Society of America*, vol. 125, no. 1, pp. 360–371, 2009.

[99] Van den Bogaert, T., Klasen, T. J., Moonen, M., Van Deun, L., and Wouters, J., "Horizontal localization with bilateral hearing aids: Without is better than with," *The Journal of the Acoustical Society of America*, vol. 119, no. 1, pp. 515–526, 2006.

[100] Wang, D., "Time-frequency masking for speech separation and its potential for hearing aid design," *Trends in Amplification*, vol. 12, no. 4, 2008.

[101] Wehr, S., Zourub, M., Aichner, R., and Kellermann, W., "Post-processing for BSS algorithms to recover spatial cues," in *Proc. International Workshop on Acoustic Echo and Noise Control (IWAENC)*, (Paris, France), 2006.

[102] Wittkop, T. and Hohmann, V., "Strategy-selective noise reduction for binaural digital hearing aids," *Speech Communication*, vol. 39, no. 1-2, pp. 111 – 138, 2003.

[103] Zwicker, E. and Fastl, H., *Psychoacoustics: Facts and Models*. New York: Springer-Verlag, 2nd ed., 1999.

VITA

Jorge I. Marín was born on February 22, 1977 in Armenia, Q., Colombia. He received the Licenciado degree in Electrical Engineering and the M.S. degree in Applied Physics from Universidad del Quindío, Armenia, Colombia, in 1997 and 2004, respectively. Since 2001, he has been with the Department of Electronics Engineering at Universidad del Quindío, where he is currently an assistant professor in the areas of Digital Systems and Signal Processing, and head researcher in the Digital Signal Processing and Processors Group (GDSPROC). In 2007, he won a Fulbright-Colciencias scholarship to pursue a Ph.D. degree in Electrical and Computer Engineering at the Georgia Institute of Technology. While his Ph.D. studies he worked as an intern for National Semiconductor Corporation (now Texas Instruments) at Santa Clara, California in Summer 2009 and 2010, and Starkey Laboratories at Eden Prairie, Minnesota, in Summer 2011, working in projects related to noise-reduction and noise-cancellation methods for industrial applications and hearing aids. His research interests include signal processing algorithms, DSP hardware systems, and hearing aids.